Data-Driven Investment Risk Assessment for BESS

Integrating Large-Language-Models through a Design Science Research Approach

SEN2331: CoSEM Master Thesis

J.Q. van Hoeve

Delft University of Technology

TU Delft

Data-Driven Investment Risk Assessment for BESS

Integrating Large-Language-Models through a Design Science Research Approach

by

J.Q. van Hoeve

In partial fulfilment of the requirements for the degree of:

Master of Science in Complex Systems Engineering and Management

at the Delft University of Technology,
to be defended publicly on 31/08/2026 at 15:00.

First supervisor:	Prof.dr. M.E. Warnier	TU Delft, Multi Actor Systems
Second supervisor:	Dr. Y. Ding	TU Delft, ICT
Company supervisor:	D. Heijning, MSc	GIGA Storage B.V.
Project Duration:	February 2026 - August 2026	
Faculty:	Technology, Policy & Management	

Cover: Lithium-ion Energy Storage project Buffalo, Lelystad, The Netherlands, by GIGA Storage B.V.

TU Delft

Preface

Dear reader,

I never expected to learn so much from writing a master's thesis. It has been a splendid journey, and one I feel fortunate to have taken. During this time I had the honour of speaking with a great many experts who work, every day, on the energy infrastructure that makes the energy transition possible. Being able to ride along with GIGA Storage at this particular moment in its growth elevated the level of this thesis in a way that I could not have achieved on my own. Working and learning in this environment was genuinely fun as well as instructive, and I am convinced that the experience will shape the rest of my career.

I owe sincere gratitude to the people who helped me bring this work together. First, to my first supervisor, Martijn Warnier, who gave me exactly the right amount of guidance and steered my methodological choices. To my mind, your supervision turned this thesis into a proper interaction between scientific research and industry contribution. Second, to Aaron Ding, whose expertise helped me see the gaps in my thinking about large language models that I would never have found alone, and who thereby strengthened the scientific grounding of the work considerably.

I also want to thank the entire team at GIGA Storage, who offered me many relevant discussions and such open insight into their processes. Those conversations will stay with me for a long time. I am equally grateful to everyone, within the organisation and beyond it, who made time for an interview and shared their expertise so openly, and to the colleagues who were always willing to think along on thesis topics, not least during the everyday lunches. In particular I want to thank Dignus, for patiently answering the endless stream of questions I brought to him.

I believe that, with the right checks and balances between an LLM and its user, this transition can advance: by working more efficiently and in a more data-driven way, and by choosing more wisely which projects are worth the energy we put into them. I hope my prototype can be a starting point for exactly that, and that the sector can find an even greater boost in realising utility-scale battery storage. Above all, I am grateful for the knowledge I gained along the way, which gives me a thorough backbone to keep working in the energy transition.

I hope you enjoy reading this thesis as much as I enjoyed making it.

J.Q. van Hoeve
Delft, August 2026

Executive Summary

The energy transition depends on storage. Battery Energy Storage Systems (BESS) are being deployed at record pace, and investors compete to acquire the projects behind that growth. Before any acquisition comes screening: an investor receives a short teaser, later a data room full of documents, and must decide within weeks whether a prospect is worth pursuing. This thesis studies that screening step at GIGA Storage B.V. (hereafter: the BESS developer), a Dutch company that develops, acquires, owns, and operates utility-scale BESS, and asks:

How can a Large Language Model-based screening instrument be designed to convert unstructured prospect documentation into a screening record that supports early-stage investment risk assessment for utility-scale BESS acquisition?

The problem underlying this question is first established empirically. Interviews with nine experts and an analysis of the acquisition process reveal five recurring bottlenecks in current practice. Screening relies heavily on information supplied by the selling party (the prospect developer). Several specialists read the same documents in parallel without a shared overview. There is no cross-functional evidence gate that forces a structured go/no-go moment. Assumptions shift between phases and between advisors without being reconciled, and the evaluation is directed mainly at the current state of a project rather than at the state it must reach to become operable and profitable. The literature offers no ready method for these conditions: established techniques need historical data or quantified inputs, and early-stage BESS screening has neither.

Following Design Science Research, the thesis converts this problem into four results that build on one another. A structured literature review, merged with the interview findings, delivers a catalogue of 52 investment risk factors across 12 categories, mapped onto the 4P screening structure used in practice (Plot, Power, Permit, Profit); to the author's knowledge, it is the first factor-level risk catalogue for utility-scale BESS acquisition. Requirements elicitation delivers 41 requirements, of which seven define the design core. A conceptual process design then splits the work: assessing a prospect requires judgement, but most screening effort goes into something more tractable, turning unstructured prospect documents into structured facts. The design decomposes LLM-supported prospect analysis into four blocks that converge on the assessment, and this thesis realises the foundational block, filling the checklist, as a working prototype.

The prototype reads a teaser and a data room and fills a 152-item checklist in three phases. Every answer carries an evidence state (found, probably found, or not found), a source with page number, and a quote that is checked mechanically against the cited document. A human reviewer gates every phase: feedback and approval can be given on the extraction, critical values must be confirmed explicitly, and once a person edits or confirms an item, no later model output silently overwrites it. In the final phase, the tool verifies key claims against publicly available sources such as company registers and grid operators.

The frozen prototype was then evaluated on nine real prospects from the acquisition pipeline. A first quick extraction costs a few cents and completes within a minute; a complete deep run costs a few dollars. The public-source check corrected or extended 62% of the claims it verified, confirming the value of an independent leg in a process that leans on seller-provided information. Three near-identical sibling prospects enabled a consistency probe: 87% of items were classified identically, and the divergence traced mainly to gaps in the item specification rather than to the model. A group session with six practitioners endorsed the design and its scope, and the team asked to start using it. Two record walk-throughs with experts who know one acquired prospect first-hand completed the picture. In the first, the expert judged the record a usable screening basis and identified one substantive extraction error, which was corrected at the review gate. The second, an in-depth due-diligence audit, delivered the evaluation's sharpest finding: the extraction was frequently right in substance but drew on the seller's own overview documents rather than the underlying evidence (61% of quotations, against the 5 to 10%

that the auditing expert considers adequate). That shortfall lies in the item specification and in the routing of documents to the model, both properties of the design rather than of the model, and both improvable by the same means this thesis demonstrates.

The result is therefore a tested proof of concept, which provides the foundation on which LLM-supported investment risk assessment can be built. The structuring layer proved technically feasible under the tested configuration, received positive formative feedback from the practitioners who would use it, and is governed by checks and balances that keep the human in control: evidence states that prefer an honest absence, verified sources behind every claim, human decisions that cannot be silently overwritten, spending capped in code, and a model that may not claim checks it did not perform. Two quantities that an adoption decision turns on remain unmeasured, and they define the follow-up study: extraction accuracy and time saved against a known baseline. Beyond the prototype, the thesis contributes the risk-factor catalogue, the requirements specification, a fully documented and reproducible prompt set, and eight design principles for LLM integration in confidential high-stakes document work.

Screening more prospects, more rigorously, with the same expert capacity means that viable storage projects are identified and funded sooner. This thesis contributes one component of that development: a working and formatively evaluated instrument that applies LLMs to the screening step, in support of a more efficient realisation of resilient electricity infrastructure.

Contents

Preface i

Executive Summary ii

Nomenclature ix

1 Introduction 1
- 1.1 Academic and Practical Relevance . . . 2
- 1.2 Research Objective . . . 2
- 1.3 Thesis Structure . . . 3

2 Foundations: BESS Investment Risk and Assessment Methods 4
- 2.1 BESS in the Energy Transition . . . 4
- 2.2 BESS Project Development Process . . . 5
- 2.3 Investment Risk in Infrastructure Development . . . 6
- 2.4 Methods for Assessing Infrastructure Investment Risk . . . 6
- 2.5 Risk Assessment by Large Language Models . . . 8
- 2.6 Knowledge Gap . . . 10

3 Design Science Research Design 11
- 3.1 Research Questions . . . 11
- 3.2 Research Paradigm: Design Science Research . . . 12
- 3.3 Research Process and Document Composition . . . 12
- 3.4 Methods of Inquiry . . . 12
  - 3.4.1 Literature review . . . 13
  - 3.4.2 Document analysis . . . 13
  - 3.4.3 Semi-structured expert interviews . . . 13
- 3.5 Methods of Design and Evaluation . . . 14
  - 3.5.1 Iterative prototyping with skills definition . . . 14
  - 3.5.2 Instrumented demonstration and consistency probing . . . 14
  - 3.5.3 Expert evaluation . . . 15
- 3.6 Research Quality, Rigour, and Ethical Considerations . . . 15

4 Current Practice at a BESS Developer: Stakeholders, Process, and Bottlenecks 17
- 4.1 Stakeholders in BESS Development . . . 17
- 4.2 Current Investment Screening Process . . . 19
- 4.3 Five Recurring Bottlenecks in Current Practice . . . 21
- 4.4 From Bottlenecks to Requirements . . . 22

5 Investment Risk Factors for BESS Acquisition 23
- 5.1 Constructing the Risk-Factor Catalogue . . . 23
- 5.2 The Risk-Factor Catalogue . . . 25
- 5.3 Limitations of the Catalogue . . . 27
- 5.4 From the Catalogue to a Prototype . . . 27

6 Requirements for Data-Driven Decision Support 28
- 6.1 Methodology for Requirements Elicitation . . . 28
- 6.2 Classification Framework . . . 28
- 6.3 Prioritisation using MoSCoW . . . 29
- 6.4 Five Requirement Categories . . . 30
- 6.5 Synthesis . . . 32

7 Artefact Design and Development **33**
7.1 A Four-Block Process for LLM-Integrated Prospect Analysis . . . 33
7.2 Data Structure and Indicators . . . 35
7.3 Decision Logic for a Three-Phase Pipeline . . . 36
7.4 Implementing a Structuring Prospect-Data Prototype . . . 37

8 Demonstration and Evaluation **41**
8.1 Evaluation Approach and Evidence Base . . . 41
8.2 Results of the Demonstration . . . 42
8.3 Consistency Under Near-Identical Inputs . . . 43
8.4 Expert Evaluation . . . 44
8.5 Synthesis: Coverage and Boundaries . . . 47

9 Discussion **48**
9.1 Four Positions on the Findings . . . 48
9.2 Reflection on the Research Design . . . 49
9.3 Contribution as Prescriptive Knowledge . . . 50
9.4 Practical Implications . . . 51
9.5 Public API or In-House Model? . . . 52
9.6 Generalisability . . . 53
9.7 Limitations . . . 53
9.8 Future Research . . . 54

10 Conclusion **55**

References **60**

A AI Statement **68**

B Interview Materials **70**
B.1 Overview of the Conducted Interviews . . . 70
B.2 Interview Protocol . . . 71
B.3 Conducted Interview Summaries . . . 72

C Expert Evaluation Sessions **79**
C.1 Group Feedback Session . . . 79
C.2 Record Walk-Through, Business Development . . . 81
C.3 Record Walk-Through, Legal and Due Diligence . . . 81

D Prototype Transparency **88**
D.1 The Checklist Instrument . . . 88
D.2 The Skill Files . . . 90
D.3 API Payloads, Call Trace, and Cost . . . 96
D.4 Illustrated Prototype Demonstration . . . 97
D.5 Development Log . . . 106
D.6 From Prototype to Production: Security and Data Handling . . . 108

E The LLM-Integrated Analysis Process **109**

F Full Risk-factor catalogue **114**

G Full Requirements Specification **118**

List of Figures

2.1 Extended version of Hameed and Træholt [36] BESS project development lifecycle. . . 5

3.1 Research-flow diagram: the five DSR activities of Johannesson and Perjons [53]. 13

4.1 Power-interest grid of stakeholders in early-stage BESS investment screening. 18
4.2 The BESS developer's project lifecycle. 19
4.3 Multi-actor process model of the project acquisition workflow. 20

5.1 Flow of risk-factor identification and selection. 24

7.1 Master process of LLM-integrated prospect analysis. 34
7.2 The filling-the-checklist pipeline. 36

9.1 The three cycles of Hevner [39], instantiated for this research. 50
9.2 Positioning LLM applications in screening and development effort. 52

D.1 Creating the prospect: name, country and the uploaded teaser document. 98
D.2 The prospect overview while the quick extraction runs in the background. 98
D.3 After the quick extraction, the headline values are populated and the prospect is ready for review. 98
D.4 Top of the quick-data review: the map link, headline cards, the per-state summary, and the critical-values gate. 99
D.5 Confirming critical anchor values: each may be accepted, changed, or marked unknown at this stage. 100
D.6 The checklist table grouped by factor, showing findings, reasoning, evidence state, confidence, and verified source quotations. 100
D.7 Once the gates are cleared, the reviewer may approve the phase or decline it with feedback. 101
D.8 Declining the phase with written feedback, which is passed to the model on the re-run. . 102
D.9 The re-run shows the previous feedback and records the new attempt alongside the first in the per-phase results. 103
D.10 The feedback has been acted upon: the current land use is now identified, whereas it was previously absent. 103
D.11 The full-extraction review: the per-state summary, the per-phase results across attempts, and the panels for differences, critical values, and probable findings. 104
D.12 The full checklist table, with the review column on the right and the automatic quotation-verification badges in the source column. 104
D.13 Selecting a citation opens the source document inside the application at the cited page. 105

E.1 An LLM-supported assessment loop process recommendation 110
E.2 Process designs of the two recurring memory-building blocks. Top: creating the market overview. Bottom: defining the BESS assessment at the BESS developer. 113

List of Tables

2.1 Assessment of investment risk assessment methods against the four requirements. . . 8

4.1 Stakeholders in early-stage BESS investment screening. 18

5.1 The risk-factor catalogue 25

6.1 MoSCoW prioritisation criteria adapted from Agile Business Consortium [3]. 29
6.2 Functional/Performance requirements (abridged). Full descriptions in Appendix G. . . . 30
6.3 Interface requirements (abridged). 31
6.4 Quality (Non-Functional) requirements (abridged). 31
6.5 Usability/Quality-in-Use requirements (abridged). 32
6.6 Human Factors requirements (abridged). 32

7.1 The four blocks of the conceptual design and their status in this thesis. 34
7.2 Structure of the checklist instrument (version 6, 152 items). 35
7.3 Illustrative correspondence between risk factors and checklist items. 36
7.4 The three extraction phases of the implemented pipeline. 37
7.5 Four items from the filled record of the prospect Meridian. 39
7.6 Traceability from design constructs to requirements and bottlenecks. 40

8.1 Evidence strands of the evaluation and their scope of validity. 41
8.2 Demonstration campaign on the frozen prototype. 42
8.3 Divergent classifications of near-identical evidence in the consistency probe. 43
8.4 Principal findings of the practitioner evaluation. 44
8.5 Document funnel of the audited prospect Beta walk-through run. 46
8.6 Requirements where the evaluation added evidence beyond the design. 47

B.1 Overview of the conducted interviews. 70

C.1 Overview of the practitioner evaluation sessions (2026). 79
C.2 Runs executed for the walk-through, on the frozen prototype. 83
C.3 Composition of the cited sources in the second full run ($n = 75$ quotes). 84
C.4 Observed citation properties against the practitioner benchmarks. 84
C.5 Item-specification problems identified in the walk-through. 85
C.6 Classes of error observed in the walk-through. 86

D.1 Fields of one checklist item (schema version 6). 89
D.2 Composition of the instrument (152 items). 89
D.3 Extract of eight items of the checklist instrument. 89
D.4 Merge and trust rules of the master record. 90
D.5 API payload specification per call type. 96
D.6 Order-of-magnitude API cost per prospect. 97
D.7 The twenty development phases of the prototype. 106

E.1 Artefact-level limitations of the frozen prototype, their causes, and fix directions. 111

F.1 Full risk-factor catalogue 114

G.1 Functional/Performance requirements. 118
G.2 Interface requirements. 121
G.3 Quality (Non-Functional) requirements. 122

G.4 Usability/Quality-in-Use requirements. 123
G.5 Human Factors requirements. 123

Nomenclature

Abbreviations

Abbreviation	Definition
4P	Plot, Power, Permit, Profit (the BESS developer's internal screening framework)
ACM	Autoriteit Consument & Markt (Dutch regulator)
AI	Artificial Intelligence
API	Application Programming Interface
BBN	Bayesian Belief Network
BESS	Battery Energy Storage System
BPMN	Business Process Model and Notation
DD	Due Diligence
DSO	Distribution System Operator
DSR	Design Science Research
EPC	Engineering, Procurement, and Construction
ESS	Energy Storage System
FID	Final Investment Decision
HV	High Voltage
IEEE	Institute of Electrical and Electronics Engineers
LLM	Large Language Model
M&A	Mergers and Acquisitions
MCDM	Multi-Criteria Decision Making
MCS	Monte Carlo Simulation
MoSCoW	Must have, Should have, Could have, Won't have (prioritisation method)
NBO	Non-Binding Offer
NDA	Non-Disclosure Agreement
NPV	Net Present Value
OEM	Original Equipment Manufacturer
ROA	Real Options Analysis
TSO	Transmission System Operator

Identifier conventions

Identifier	Meaning
B1–B5	Bottlenecks in current screening practice (Section 4.3)
E1–E3	Evaluation participants new to the research (Appendix C)
P1–P9	Interview participants (roster in Appendix B)
Alpha–Iota	Anonymised code names of the nine demonstration prospects (Chapter 8)
Meridian	The fictional prospect created for the illustrated demonstration (Appendix D.4)
QL-, FR-, IF-, US-, HU-	Requirement categories: Quality, Functional/Performance, Interface, Usability, Human Factors (Chapter 6)
SELL-, PLOT-, POWER-, PERMIT-, PROFIT-	Checklist item prefixes per factor (Chapter 7, Appendix D.1)

Definitions

Term	Meaning in this thesis
Prospect	A candidate BESS project offered for acquisition and under screening.
Prospect developer	The party that develops a project opportunity and offers it for acquisition after securing, or creating a credible pathway to, site control, permitting and a grid connection; the selling party in the transaction.
BESS developer	A company that develops utility-scale BESS projects through to construction, ownership, and operation, whether greenfield or through acquisition. The commissioning organisation of this research is one.
Teaser	The short introductory document(s) a seller shares before access to the data room is granted.
Data room	The (virtual) collection of project documents a seller opens for due diligence.
Checklist item	One of the 152 defined information items that together describe a prospect.
Evidence state	The classification of an item's answer as found, probably found, or not found.
Critical item	A checklist item whose value must always be confirmed by a human reviewer.
Anchor value	A confirmed critical value (for example the seller name or the size in MW) carried into later phases to target searches.
Skill	A versioned, editable instruction file that defines one LLM task (the design object).
Prompt	The assembled instruction text of one specific API request (the runtime object), of which the skill is the reusable source.
Phase	One of the three extraction stages (quick, full, gaps), each ending in a human approval gate.
Conflict	A recorded disagreement between findings for the same item, for example between two phases or between a phase and a user edit.
Master record	The merged, living checklist per prospect that combines all phase results and every human decision.
Walk-through	The evaluation format in which a domain expert reviews a filled prospect record item by item with the researcher (Chapter 8, Appendix C).
Artefact	The designed output of Design Science Research: here, the structured screening method together with its working instantiation. Referred to throughout this thesis as the prototype (Section 3.2).

1

Introduction

Electricity systems are reorganising around variable renewable generation, and this reorganisation makes flexibility a scarce and valuable system resource. Utility-scale BESS have become the leading grid-scale answer to this scarcity, because they can absorb and release energy within seconds and can be built almost anywhere on the network [89, 108].

The market has responded rapidly, with global battery storage additions reaching approximately 108 GW in 2025, roughly 40% more than the year before, and around four-fifths of this new capacity was utility-scale [46]. Europe followed the same pattern: 36 GWh of new battery capacity was installed in 2025, and for the first time utility-scale systems accounted for the majority of annual additions [87]. Utility-scale BESS has thereby developed from a niche technology into an established infrastructure asset class that attracts substantial private and institutional capital [81].

The time-to-market of a battery project is determined primarily by permitting, grid connection and financing rather than by construction, and regulatory uncertainty around these steps is identified as a key brake on further growth [46]. The Netherlands illustrates this constraint in an acute form. Structural grid congestion has placed an estimated 12,000 to 14,000 companies on waiting lists for a new or expanded grid connection [95]. Under these conditions, a project site with a credible route to a grid connection and a permit is a scarce commodity, and competition for such sites begins long before anything is built.

This scarcity has shaped a distinct market structure. Prospect developers (the parties that originate early-stage projects and later offer them for sale) identify candidate sites, secure initial positions on land, grid access, and municipal support, and then offer the resulting development projects to specialised acquirers. For the acquiring party, the decision to buy is taken at the point in the project lifecycle where the least is known. The grid connection is often still pending, the permits have not yet been granted, the land position may be only partially secured, and the revenue model rests on assumptions rather than contracts, while critical estimates depend on external actors (the grid operator, the permitting authorities) that cannot yet provide certainty at this stage. Most of the information available to the acquirer originates, moreover, from the party selling the project. Errors are costly in both directions: acquiring a project that later proves unviable turns development capital into sunk cost, while declining a project that would have succeeded forgoes both the private return and the storage capacity that the electricity system requires.

The rapid growth of the asset class is thereby outpacing the methods available to assess projects reliably in the early investment stages. In current practice, these acquisition judgements are made through largely manual screening: experts read fragmented project documentation, form a view per discipline, and combine their views into a go/no-go decision, with external technical and legal due-diligence advisors engaged only in the later stages of a transaction. This way of working is knowledge-intensive and hardly standardised, which makes it difficult to scale when the number of candidate projects grows. How this problem manifests in a real acquisition workflow is examined empirically in Chapter 4.

At the same time, advances in artificial intelligence are changing what is technically possible in exactly this kind of document-heavy assessment work. AI methods are being applied at a growing pace across the energy sector [96, 101]. In particular, Large Language Models (LLMs) can operate directly on the heterogeneous, largely textual material that characterises early-stage project documentation. Whether and how this capability can be used responsibly to support investment risk assessment in BESS acquisition is, however, an open question. Chapter 2 examines what established assessment methods and the emerging LLM literature can and cannot offer in this setting.

1.1. Academic and Practical Relevance

Academically, this thesis is positioned at the intersection of two bodies of knowledge that have so far developed largely in parallel: the literature on investment risk assessment for energy infrastructure, and the literature on AI-based decision support. The early-stage BESS acquisition context combines properties that make it a demanding test case for both. Information is scarce and largely qualitative, the pool of comparable completed projects is small, and screening must be performed quickly across multiple candidates. Chapter 2 substantiates the claim that existing assessment methods and the published evidence on LLMs leave this specific setting unaddressed, and derives the knowledge gap from it. The intended academic contribution is design knowledge in the Design Science Research (DSR) sense [40]: a screening method grounded in an explicit problem analysis, a working prototype that operationalises it, and an evaluation of both against current practice. To the author's knowledge, it is among the first structured explorations of how LLMs can support this early screening step, and the instrument is designed to prepare the material on which experts base their judgement, while the judgement itself remains with them.

Practically, the research addresses a capability that acquirers in this asset class currently lack. A structured, data-driven screening instrument can make assessments more consistent across projects and evaluators, shorten the time from receiving an opportunity to a first substantiated view, preserve the reasoning behind past decisions as institutional knowledge, and allow a small team to screen a larger pipeline without a proportional increase in expert effort. Better screening also has a societal dimension. The pace at which storage capacity is realised depends in part on how quickly and how well viable projects are identified and funded. AAn instrument that reduces both wrongful acquisitions and wrongful rejections may therefore contribute to the deployment of the flexibility that congested electricity systems require, although the size of that contribution is not quantified in this thesis. Because the research is conducted inside a real acquisition environment, its findings are directly relevant to the host organisation and, with appropriate caution, to comparable BESS developers operating in the Netherlands, Belgium, and Germany.

1.2. Research Objective

The objective of this research is to design and evaluate an LLM-based decision support prototype for early-stage investment risk assessment in utility-scale BESS acquisition. The prototype targets the foundation on which any such assessment rests: converting the unstructured documentation of a prospect into a structured, source-attributed, and verifiable screening record, under explicit human control. It does not itself produce the investment-risk judgement, which remains with the expert. Around that working core, the research delivers the substantive and behavioural specifications a full assessment requires (the risk factors to cover and the requirements to satisfy) and the process design in which the prototype is the first building block.

The research follows the DSR methodology, in which a designed solution is built to address a relevant real-world problem while generating scientific knowledge about that solution and its use [53]. To ground the design in a realistic decision environment, the research is conducted in collaboration with GIGA Storage B.V. (hereafter: the BESS developer), a Dutch company that develops, acquires, owns, and operates utility-scale BESS in the Netherlands, Belgium, and Germany. The BESS developer acts as the commissioning organisation and provides access to the screening materials, decision routines, and domain experts that a study of this kind requires. The case context is described in Chapter 4.

The main research question and its sub-questions are deliberately not stated in this chapter. They follow directly from the knowledge gap established in Chapter 2 and are therefore formulated at the

start of Chapter 3. The objective stated above guides that derivation.

The scope of the research is bounded in three ways. First, it concerns the pre-acquisition screening phase, from the moment an investment opportunity is received to the final investment decision; post-acquisition development, procurement, and operations fall outside the scope. Second, it takes the perspective of the acquiring BESS developer rather than that of the prospect developer, the grid operator, or the permitting authority. Third, its empirical and regulatory grounding lies in the Netherlands, Belgium, and Germany, the jurisdictions in which the BESS developer operates.

1.3. Thesis Structure

The remainder of this thesis follows the DSR activity chain from problem to evaluated prototype. Chapter 2 develops the theoretical foundations: it situates BESS investment in its technical and economic context, conceptualises early-stage investment risk, reviews the principal families of risk assessment methods, examines the evidence on LLM-based risk assessment, and derives the knowledge gap. Chapter 3 translates this gap into the research questions and the methodological design. Chapter 4 explicates the problem empirically by analysing the stakeholders, the screening process, and the recurring bottlenecks at the BESS developer. Chapter 5 constructs the catalogue of investment risk factors that an assessment must cover, and Chapter 6 specifies the requirements that the prototype must satisfy; together these two chapters define what must be assessed and how the assessment must behave. Chapter 7 presents the design and development of the prototype. Chapter 8 demonstrates the prototype and reports its evaluation by domain experts against current practice. Chapter 9 discusses the positions this thesis takes on its findings, the practical implications, the generalisability, and the limitations. Chapter 10 answers the research question and states the academic and societal contributions.

2

Foundations: BESS Investment Risk and Assessment Methods

This chapter develops the theoretical and methodological foundations on which the remainder of the thesis is built, working from the general to the specific. Section 2.1 situates BESS within the energy transition and identifies the sources of uncertainty that characterise their development. Section 2.2 reviews how BESS projects are developed in practice and how information accumulates across their decision gates. Section 2.3 conceptualises investment risk in this context. Section 2.4 critically reviews the principal families of methods applied to infrastructure investment risk assessment and derives the requirements a suitable approach must satisfy. Section 2.5 examines what the published evidence shows about LLMs for risk assessment in adjacent domains, and Section 2.6 states the knowledge gap that motivates the research design of Chapter 3.

2.1. BESS in the Energy Transition

The decarbonisation of electricity systems around variable renewable generation places flexibility at the centre of infrastructure planning, and BESS have emerged as a leading response, combining rapid response capability with steadily declining costs. The IPCC Sixth Assessment Report identifies utility-scale storage as an enabling technology for deep decarbonisation, and the International Renewable Energy Agency projects that global installed storage capacity must increase more than sixfold by 2030 relative to 2022 levels to remain consistent with a 1.5°C pathway [45, 49]. At the European level, the revised Renewable Energy Directive's binding 2030 target of at least 42.5% renewables in final energy consumption implies a commensurate increase in balancing demand [23]. Energy storage addresses the structural mismatch between variable generation and demand by decoupling production from consumption, and at grid scale it can additionally defer costly network reinforcement [64, 89, 108]. Among available storage technologies, lithium-ion BESS have grown fastest, driven by high round-trip efficiency (85–95%), millisecond-scale response, and continuous capital cost reductions [26, 56].

The economic case for utility-scale BESS rests on value stacking: energy arbitrage alone is insufficient under current cost structures, but combined with ancillary services such as Frequency Containment Reserve (FCR) and automatic Frequency Restoration Reserve (aFRR) it substantially improves returns [41, 105, 81]. The feasibility of value stacking depends on two infrastructure conditions. The regulatory framework governing market participation and revenue stacking varies across European jurisdictions and remains subject to revision, introducing policy uncertainty that can materially affect bankability [86, 31]. And grid connection quality directly constrains revenue realisation: non-firm connections may accelerate deployment, but they introduce curtailment risk that erodes the revenue base [99]. Battery degradation adds a further investment-relevant complexity: capacity loss through calendar and cycling ageing is sensitive to operational parameters that are rarely known with precision at the pre-acquisition stage [20]. Because BESS projects combine high capital costs, external grid and regulatory dependencies, and uncertain technical performance, financial modelling must be supplemented with

technical, legal, and permitting review; how such risk should be conceptualised is addressed after the development process itself.

2.2. BESS Project Development Process

The development of utility-scale BESS projects is a multi-stage, iterative process characterised by the gradual reduction of uncertainty through successive decision gates. At the outset, projects carry a high volume of unknowns, and the initial capital expended is risk capital subject to total loss if the project proves unviable; as development progresses, information improves, allowing developers to identify fatal flaws early and abandon unworkable projects before high-capital commitments [88].

The dominant technical framing in the BESS literature is the Engineering, Procurement, and Construction (EPC) framework of Hameed and Trætholt [36], which structures development into five sequential stages: feasibility assessment, permitting and regulatory approval, procurement and financial close, construction and installation, and commissioning and handover. The first stage establishes the foundation of the business case through site screening on land availability, infrastructure proximity, and environmental constraints [36, 74]. The second is often the most critical of early-stage development, since it requires securing a formal grid connection offer and legal authorisation to construct [36, 74]. Grid-connection choices are made early and couple several variables at once: developers must select the voltage level and physical connection point together with the requested capacity, and because connection costs rise stepwise with voltage level, there is an economic optimum between the voltage level and the size of the connection for a given project [37, 36]. Recent work adds the trade-off between firm and non-firm arrangements: firm connections offer uninterrupted access at long waiting times and high fees, whereas non-firm arrangements provide faster integration at the cost of curtailment uncertainty [99]. The third stage marks the transition from development risk capital to project finance [88], and the final stages bring the design to operational reality through construction and Site Acceptance Tests [36, 94].

Other frameworks describe the same process from different vantage points. Nordic Energy Research and Nordic Council of Ministers [74] distinguish project development from implementation, emphasising that financial institutions engage only once a project reaches the bankable feasibility milestone. The World Bank defines the process through contract-management stages, prioritising identification, screening, and option selection [94]. Springer [88] introduces the SROPTTC framework (Site, Resource, Off-take, Permits, Technology, Team, Capital) and treats development as iterative belief updating, in which the project's value is recalculated after each incremental investment in data gathering. The Storage Readiness Level framework adapted from CO_2 storage practice grades site-specific maturity on a nine-level scale governed by sequential decision gates [4]. Two extensions complete the investment horizon: operation constitutes a distinct, value-critical phase in which assets may be refinanced or traded on realised cash flows [88, 78], and the limited technical lifetime of battery systems introduces end-of-life considerations that early-stage decisions must already anticipate [94]. Figure 2.1 integrates these perspectives into an extended view of the BESS project lifecycle, with the EPC core highlighted as the early- to mid-stage focus of this thesis.

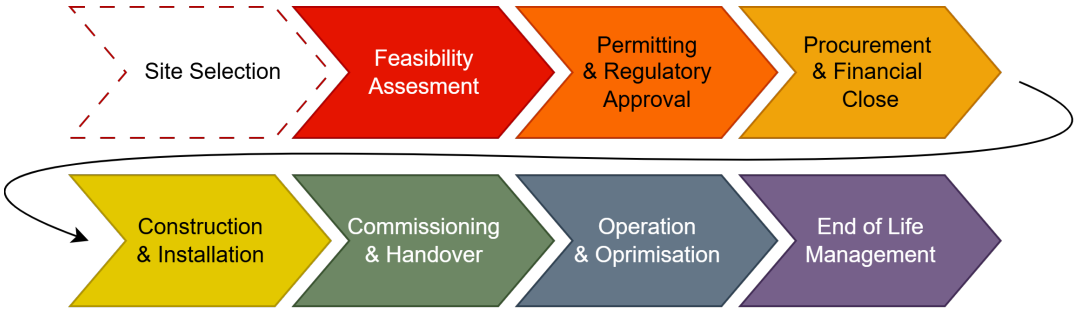

Figure 2.1: Extended version of Hameed and Trætholt [36] BESS project development lifecycle.

In these frameworks, information accumulates gradually across the development stages. Springer [88] describes development as a process of belief updating, in which broad desktop studies give way to increasingly investment-grade data. The SROPTTC elements are progressively assembled to support

advancement. In practice these information streams are rarely integrated from the start; they are distributed across technical, legal, financial, and regulatory domains, each requiring manual interpretation [107, 101]. Rasmussen and Niles [78] therefore argue for a pull-based logic in which the decision requirements of the next gate determine which information is necessary and sufficient. For early-stage screening this means that decision quality depends not only on what information is available, but on how effectively it is structured and linked to the decision at hand, a problem to which the assessment methods of Section 2.4 respond in different ways.

2.3. Investment Risk in Infrastructure Development

Investment risk in infrastructure is broadly defined as the effect of uncertainty on an organisation's ability to meet its strategic objectives, encompassing not only the probability of adverse outcomes but also the conditions under which decisions must be made [1, 17]. Two structural features are particularly consequential for BESS acquisition.

The first is the distinction between systematic and idiosyncratic risk [22, 25]. Systematic risks, such as electricity price volatility, policy shifts, or changes in ancillary market design, affect all projects in a portfolio and cannot be diversified away [25, 35]. Idiosyncratic risks are project-specific and arise from information asymmetries between the acquirer and the prospect developer [22, 102]: the quality of a grid connection offer, the legal clarity of land titles, and the robustness of permitting prospects are conditions the seller holds privately and the acquirer must verify. In early-stage BESS screening, idiosyncratic risks dominate, because the macro-economic environment is shared across projects while each site's feasibility conditions are unique and largely unverified.

The second is irreversibility. Dixit and Pindyck establish that when investments involve significant sunk costs, the option to wait and gather additional information before committing capital has genuine economic value that standard Net Present Value analysis does not capture [24]. This motivates rigorous front-end screening: reducing idiosyncratic uncertainty before the irreversible commitment preserves the acquirer's options to defer, renegotiate, or abandon, whose value is highest precisely when uncertainty is greatest.

A further complication is that infrastructure projects systematically underperform their pre-investment forecasts. Flyvbjerg documents persistent cost overruns and benefit shortfalls across sectors, attributing them to optimism bias and to strategic misrepresentation by parties with an interest in approval [32, 33]. The recommended corrective is reference class forecasting: grounding assessments in the empirical distribution of outcomes for comparable completed projects [33]. In the BESS acquisition context, this implies treating seller-provided information with systematic scepticism and benchmarking against observed outcomes wherever comparable data exists.

Finally, the literature offers risk typologies relevant to structuring BESS screening. Miller and Lessard distinguish market, completion, and institutional risks, and find institutional risks the most frequent cause of project failure because they involve actors whose decision logic is incommensurable with the investor's [68]; this maps directly onto the BESS context, where grid operators and permitting authorities operate under timescales and mandates independent of commercial schedules. Abba et al. complement this with a holistic framework identifying technical, economic, social, and institutional categories that interact rather than operating independently [1]. Early-stage BESS screening is thus dominated by project-specific risks: prospect developers may provide optimistic information, grid and permitting outcomes depend on external institutions, and acquisition commits capital that may not be recoverable. How these risks can be assessed in a structured and data-driven way under fragmented information is taken up in Section 2.4.

2.4. Methods for Assessing Infrastructure Investment Risk

The previous section established that early-stage infrastructure investment is characterised by high uncertainty, information scarcity, and staged decision-making. This section reviews the principal families of methods applied to energy infrastructure risk assessment (qualitative expert judgement, Multi-Criteria Decision-Making (MCDM), Monte Carlo Simulation (MCS), Real Options Analysis (ROA), and Bayesian Belief Networks (BBNs)) and evaluates each against the constraints that early-stage BESS screening imposes.

Expert Judgement and Qualitative Scorecards

The most widely used approach in practice is structured expert judgement, operationalised through checklists, red-flag reports, or scoring sheets [32, 11]. It is highly flexible, accommodates the narrative information that characterises pre-acquisition DD, and is positioned in the M&A literature as an essential quality-control mechanism against information asymmetry [102, 11]. Its limitations are equally documented. Individual assessments are subjective and difficult to replicate, reducing consistency across projects and evaluators [1, 51]. Qualitative scores treat risk dimensions as independent, whereas BESS failure modes involve cascading interdependencies [55]. And expert-based processes do not scale: each assessment requires specialist time that is difficult to standardise [72, 11].

Multi-Criteria Decision-Making

MCDM methods formalise the weighting and aggregation of heterogeneous criteria, and approaches such as AHP, Best-Worst and fuzzy TODIM have been widely applied to renewable energy investment [109, 38, 62]. Their strengths are explicit, auditable weighting and, in fuzzy extensions, linguistic scoring for vague inputs [1, 110, 54]. Two structural limitations constrain their use here: the models are static, assigning weights at a fixed point rather than following the sequential arrival of information across gates [109, 50], and standard aggregation assumes criteria independence, which the interacting grid, permit, and land dimensions of BESS acquisition violate [1, 55]. Weight elicitation through expert panels also reintroduces subjectivity and limits scalability [109, 38].

Monte Carlo Simulation

MCS introduces probability distributions over uncertain inputs and produces distributions of outcomes such as NPV or IRR [84, 30, 16], which suits the stochastic revenue streams of BESS [63, 41]. Its preconditions are not met at the screening stage. The most consequential early-stage factors (a permitting outcome, the firmness of a connection offer, the validity of a land title) are categorical and resist statistical characterisation, so imposed distributions create false precision [1]. Standard MCS also assumes input independence, while BESS risks co-vary [16, 1], and the method values projects that are already defined rather than screening ones whose feasibility is unresolved [84].

Real Options Analysis

ROA recognises the economic value of managerial flexibility, treating commitments as options that can be deferred, staged, or abandoned [63, 70]; for BESS, Locatelli, Invernizzi, and Mancini [63] show that ROA produces materially different valuations than deterministic NPV when volatility and flexibility are high. It is nevertheless rarely deployable in early-stage screening. Its formulations require the underlying uncertainty to be reducible to a tractable stochastic process, presupposing analytical precision that pre-acquisition screening lacks [70, 57], and it evaluates well-defined opportunities with known exercise prices, which early-stage prospects do not offer [5, 63]. ROA is therefore most useful near financial close rather than as a first-pass screen [70].

Bayesian Belief Networks

BBNs encode uncertain variables and their conditional dependencies in a directed acyclic graph, supporting causal propagation of risk and belief updating as evidence arrives [51, 55]. Islam et al. [51] demonstrate their power for interdependent cost-overrun risks in power plants, and Khan et al. [55] extend them to socio-environmental risk. Their strengths align well with the epistemic structure of early-stage risk: they accept mixed quantitative and expert-elicited inputs and support the gate-structured logic of development [51, 1]. The practical constraint is construction cost: eliciting conditional probability tables is demanding, degrades in reliability without observed data for validation, and must be substantially revised per project type and jurisdiction [51, 55, 1].

Synthesis: Requirements for a Suitable Assessment Approach

Each method performs well in the conditions for which it was designed, and each encounters a specific obstacle in early-stage BESS screening. Reading across the evidence, four requirements emerge that any suitable approach must satisfy. Table 2.1 summarises the assessment.

Requirement 1: Operability under information scarcity and qualitative input dominance. The most consequential early-stage factors are not numerical: permitting outcomes, the firmness of a con-

Table 2.1: Assessment of investment risk assessment methods against the four requirements.

Requirement	Expert Judgement	MCDM	MCS	ROA	BBN
R1: Qualitative / mixed-input operability	$+$	$\sim$	$-$	$-$	$+$
R2: No large historical dataset required	$+$	$+$	$-$	$-$	$\sim$
R3: Interdependency representation	$-$	$-$	$\sim$	$\sim$	$+$
R4: Scalability without per-project expert effort	$\sim$	$\sim$	$\sim$	$-$	$-$

$+$ = requirement satisfied; $\sim$ = partially satisfied or context-dependent; $-$ = requirement not satisfied.

nection offer, and the legal clarity of land titles are binary or ordinal and resist statistical characterisation [1, 32, 88]. A suitable method must accept mixed inputs without forcing artificial quantification. Expert judgement satisfies this natively [32, 11]; BBNs satisfy it through discrete states and expert-elicited priors [51]; MCDM partially, via fuzzy extensions that still require ordinal structure and a functioning panel [109, 54]; MCS and ROA do not, since both require well-specified distributions or parametric processes [84, 63].

Requirement 2: Applicability without a large historical project dataset. The pool of completed BESS acquisitions is small, recent, and heterogeneous in chemistry, duration, and connection regime [81, 19, 20], ruling out methods whose reliability depends on empirical calibration. Expert judgement and MCDM satisfy this through elicitation [11, 109]; BBNs partially, since conditional probability tables degrade without observed validation data [51, 55]; MCS and ROA do not [30, 70, 63].

Requirement 3: Capacity to represent interdependencies between risk dimensions. BESS project risks are causally linked: a delayed connection erodes the permitting timeline, which jeopardises financial close, which destabilises the revenue model [1, 51, 55]. BBNs satisfy this fully by construction [51, 55]; MCS and ROA partially, through correlated sampling and compound options [16, 63]; expert judgement and MCDM do not, the former lacking formal structure and the latter assuming independence by construction [109, 1].

Requirement 4: Scalability across multiple projects without bespoke expert effort per project. Investor-side screening assesses a pipeline under time pressure [11, 72]. A method requiring project-specific calibration (new weights, new conditional probability tables, a new stochastic model) is operationally infeasible at scale. Expert judgement and MCDM partially satisfy this (a consistent analyst or an established weight set carries across projects, at the cost of replicability and upfront elicitation [32, 11, 109, 38]); MCS partially, once a model exists; ROA and BBNs do not, since construction is the most resource-intensive step of all the methods reviewed and must be repeated per context [70, 51, 55].

Two conclusions follow from the assessment against the four requirements. The methods that are analytically most capable of representing the problem's epistemic structure (BBN, MCS, ROA) are the ones that cannot be reliably applied given the data and resource constraints of the context: their limitation lies in their fit with the decision context rather than in their analytical power. And this mismatch is compounded by market immaturity: the calibration data those methods require does not yet exist. A suitable approach must operate on scarce and largely qualitative information, require no long project history, represent interdependencies, and scale across a pipeline. LLMs warrant evaluation against exactly this profile, because they operate natively on heterogeneous textual inputs, require no task-specific training data, and can incorporate domain knowledge through structured prompting rather than statistical calibration [14]. Whether the published evidence supports their use in this setting is examined next.

2.5. Risk Assessment by Large Language Models

Section 2.4 concluded that none of the principal method families is fully calibrated to early-stage BESS acquisition. This section examines whether LLMs offer a candidate direction, guided by three questions: where AI in project development stands, what has been done with LLMs in adjacent contexts, and which limitations remain unresolved.

The broader trajectory: AI in project development and investment

The use of AI in project development is no longer a niche research topic. In a systematic review of 215 studies, Nenni et al. [72] document a sustained shift in project management research toward AI-based approaches to project risk, used to augment rather than replace expert judgement. Mertzanis [67], reviewing 178 studies on AI in investment management, reaches a structurally similar conclusion, while flagging explainability, adoption barriers, and ethics as the most consistent unresolved issues. In the energy domain, comprehensive reviews position AI as a transformative enabler across resource assessment, forecasting, grid stability, and system monitoring [101, 96], and firm-level evidence shows AI adoption significantly increasing a renewable-energy firm's ability to attract private investment [82]. What is new is the emergence of LLMs as a class of AI capable of operating natively on the heterogeneous textual material of pre-acquisition due diligence: contracts, technical reports, regulatory documents, and unstructured memoranda [52, 14].

What has been done with LLMs in adjacent contexts

The most directly relevant evidence comes from construction project risk management. Erfani and Khanjar [27] compared LLM-based classification of construction risk statements against twelve other configurations, including BERT-style approaches[1], and their substantive finding is that LLMs reach competitive performance without task-specific labelled training data, positioning them as practical tools precisely where labelled data and expert resources are scarce. Martin, James, and Chadee [66] reach a complementary conclusion on an infrastructure project in Ireland: LLM-assisted assessment substantially reduces task time and broadens the risks initially identified, but experienced practitioners judge it to lack contextual depth, and the recommended arrangement is hybrid, with the model supplying breadth and the human depth. Two of their observations carry directly into this thesis: output quality depends more on the quality of structured input context than on prompting skill, and the value is largest for less experienced practitioners, as a structured awareness tool rather than a decision aid.

In finance, Lee, Stevens, and Han [58] survey domain-adapted financial LLMs and identify a recurring pattern: general-purpose models hit a ceiling on financial tasks that domain-adapted variants exceed, particularly for information extraction from filings. Huang, Wang, and Yang [42] demonstrate this concretely, with the gap largest when training samples are small and terminology is rare, conditions that resemble BESS due-diligence material. In the energy sector, Majumder et al. [65] provide the most systematic assessment of LLM capabilities for power-system tasks, identifying four strengths (in-weight memorisation, in-context learning, tool embedding, multi-modal processing) and concluding that general-purpose LLMs perform well as decision-support co-pilots for text, document, and correlation tasks, but not as substitutes for physics-based or safety-critical computation. Taken together, this literature documents LLM applications to construction risk classification, financial-text analysis, and energy-system diagnostics; to the author's knowledge, it does not address the early-stage investment screening of utility-scale BESS projects, which combines heterogeneous pre-acquisition documents, time pressure, and consequential capital decisions in a way none of these settings reproduces.

Documented limitations

The same body of work identifies limitations directly relevant to investment risk assessment. The first is hallucination: LLMs can generate fluent but factually incorrect output, which is particularly consequential where a single misclassified risk is costly [65, 66]. The second is bias propagation: Winder, Hildebrand, and Hartmann [104] show that LLM-generated investment advice can reinforce sectoral and geographic concentration biases, increasing realised risk for users who follow it uncritically. The third is sensitivity to prompt design: small wording changes can produce materially different outputs, which constrains reproducibility and motivates explicit versioning of prompts as artefacts in their own right [27, 65]. Two further limitations are relevant here for document-heavy assessment. A model's general knowledge cannot be switched off through instruction: it may silently substitute what it knows for what the documents say, which is hazardous exactly where documents and priors diverge. And a model's use of long inputs is position-sensitive: evidence can be harder to use depending on where it appears in the context, so supplying more documents is not the same as having them read [60]. A growing engineering literature responds by making models cite their sources: generated statements

[1] Refers to computational methods for analysing and generating human language; classical NLP approaches typically require large labelled datasets for training on a specific task.

are required to carry citations, and the citations are verified against the documents they name [34, 59]. This line of work treats provenance as a property of the model's output; Chapter 9 returns to what that framing overlooks when the document set itself is assembled by an interested party.

Governing the risks of AI itself: the NIST AI Risk Management Framework

The limitations above are usually treated as technical properties of models. A complementary body of work treats them as an organisational risk-management problem: introducing an AI system into a commercial workflow creates risks that must be governed like any other operational risk. The reference framework is the Artificial Intelligence Risk Management Framework (AI RMF 1.0) of Tabassi [92]. Its starting point is that AI risk is socio-technical: harms arise not only from models and data but from how organisations deploy and use them. It defines seven characteristics of trustworthy AI (valid and reliable; safe; secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; and fair) and organises risk management into four functions: govern, map, measure, and manage. Two aspects are directly relevant here. The trustworthiness characteristics anticipate several requirements elicited in Chapter 6, from source attribution to conservative uncertainty handling. And the framework makes explicit that deploying a third-party LLM inside a confidential acquisition workflow is itself a risk decision requiring governance, not only engineering; Chapter 7 designs with these characteristics in mind, and Chapter 9 assesses which the prototype addresses.

2.6. Knowledge Gap

The preceding sections establish four findings that together specify an addressable gap.

First, early-stage utility-scale BESS acquisition is conducted under structurally distinctive conditions: information scarcity, qualitative input dominance, low project frequency for empirical calibration, interdependent risk dimensions, and time-constrained screening across multiple candidates (Sections 2.1 and 2.3). Most of these conditions are structural to the pre-acquisition stage itself: information asymmetry and idiosyncratic dominance follow from buying before certainty exists, and will not disappear as the market matures. The scarcity of calibration data is the exception: it is transitional, and it will ease as completed acquisitions accumulate, which is precisely why a suitable approach should remain valid when that data arrives rather than depend on its absence.

Second, none of the principal families of investment risk assessment methods reviewed in Section 2.4 is fully suited to these conditions. The analytically strongest methods (MCS, ROA, BBN) require distributions, tractable processes, or elicitation effort the context cannot provide; the readily deployable methods (expert judgement, MCDM) offer limited support for interdependency modelling and sequential updating. Refining an existing method does not resolve this mismatch.

Third, although LLMs have capabilities aligned to these conditions (Section 2.5), the published evidence does not address early-stage BESS investment screening. Adjacent work classifies historical construction risk statements [27, 66], processes public financial disclosure [42, 58], or supports power-system operations [65], and the documented limitations [104, 65] indicate that any LLM-based approach for this context must be designed deliberately rather than adopted off-the-shelf.

Fourth, the infrastructure-investment literature provides risk typologies at category level [68, 1] but stops short of an operationalisable set of risk indicators for BESS acquisition; indicator-level guidance exists in practitioner and grey literature but is dispersed rather than consolidated.

The combined gap is therefore threefold. At the methodological level, no structured assessment approach is calibrated to the conditions of early-stage BESS acquisition. At the substantive level, the literature does not provide the operationalised set of risk factors such an approach would assess. At the artefactual level, no LLM-based screening instrument has been designed and evaluated for this context. This thesis addresses the gap in three steps: it develops BESS-specific risk factors, converts them into prototype requirements, and implements and evaluates a screening prototype against current practice. The combined gap motivates the main research question and the sub-questions developed at the start of Chapter 3.

3

Design Science Research Design

This chapter translates the knowledge gap identified at the end of Chapter 2 into the research design of this thesis: the research questions, the Design Science Research (DSR) paradigm, the execution plan that connects questions to methods and outputs, the methods of inquiry and of design and evaluation, and the quality and ethical measures that govern them. It describes the methods and the choices behind them. What the methods produced is reported in Chapters 4 to 8, and the full records are documented in the appendices.

3.1. Research Questions

Chapter 2 concluded that early-stage utility-scale BESS acquisition is conducted under a set of structurally distinctive conditions, that none of the principal families of investment risk assessment methods is fully calibrated to those conditions, and that LLMs are a candidate direction whose applicability to this context has not yet been examined in the published literature. From this combined gap the thesis derives one main research question and five sub-questions.

Main research question

> *How can a Large Language Model-based screening instrument be designed to convert unstructured prospect documentation into a screening record that supports early-stage investment risk assessment for utility-scale BESS acquisition?*

The question is solution-oriented: it asks how a prototype with specified properties can be designed and evaluated, rather than whether a hypothesis holds about an existing state of affairs, which situates the research in the design tradition. The main question already commits to an LLM-based instrument, on the grounds established in Chapter 2; the sub-questions avoid this commitment where the work they define does not depend on it, so that the problem analysis and the requirements stand on their own.

Sub-questions

SQ1. How does early-stage investment risk assessment for BESS project acquisition currently work, and what are its recurring bottlenecks?

SQ2. Which investment risk factors are relevant for early-stage BESS acquisition, and how can they be structured for screening use?

SQ3. What requirements should a data-driven decision support prototype satisfy?

SQ4. How can the identified risk factors and requirements be realised in a working LLM-based prototype?

SQ5. How does the prototype perform under realistic screening conditions, and how do domain experts assess its fit with current practice?

The sub-questions follow the DSR logic: problem explication (SQ1), the substantive and behavioural requirements of any solution (SQ2, SQ3), realisation in a concrete artefact (SQ4), and evaluation of

that artefact (SQ5). SQ4 covers two layers that are realised differently. The requirements are realised in the implemented prototype, while the risk factors are realised as the specification of the judgement layer that the prototype prepares but deliberately does not perform (Section 7.1).

3.2. Research Paradigm: Design Science Research

The methodology must align with the nature of the research problem and the type of knowledge to be generated [21]. Because the main question is solution-oriented and the intended contribution is a novel artefact that addresses a real-world problem while generating scientific knowledge about it [40], the thesis adopted the DSR framework of Johannesson and Perjons [53], organised around five core activities: explicating the problem, defining requirements, designing and developing the artefact, demonstrating it, and evaluating it. The activities are iterative rather than sequential: insights from demonstration and evaluation feed back into design.

As the evaluative lens over these activities, the thesis uses the three-cycle view of Hevner [39]: the relevance cycle draws problem and requirements from practice and returns the artefact to it for evaluation, the rigour cycle connects the research to the knowledge base of Chapter 2, and the design cycle alternates between building and evaluating. The cycles are instantiated for this research, with a diagram, in Section 9.2.

Two known limitations of DSR shape how this thesis proceeds. Artefacts are developed within a specific organisational context, which constrains direct generalisation, so the thesis follows the strategy of Iivari [44]: solve a specific problem in a local practice, here the screening process at a BESS developer, then distil prescriptive knowledge that can inform a more general solution (Chapter 9). DSR artefacts are also often reported with insufficient design rationale to be transferable, leaving a "black box" prototype [71], a critique that is acute where the artefact contains an LLM whose internal reasoning cannot be traced; the operational response is set out in Section 3.6.

The artefact combines two of the artefact types distinguished in design science [53]: a *method*, the structured screening procedure organised around defined investment risk factors, and an *instantiation*, the working prototype that operationalises the method through an LLM-supported screening workflow. Because the method can be specified independently of the particular LLM, the prescriptive contribution of this thesis outlasts the specific models used to build the prototype. Throughout this thesis, this combined artefact is referred to simply as the prototype. Stated in one sentence, it is a human-gated LLM pipeline that converts prospect documents into a structured, source-attributed screening record; it does not perform the final investment-risk judgement.

3.3. Research Process and Document Composition

The execution plan distributes the five DSR activities over the chapters of this thesis: each chapter answers one sub-question through a defined combination of inputs, methods, and outputs, and every chapter's output is used as input by at least one downstream chapter, which is the internal consistency check the plan is designed to pass [106].

Figure 3.1 presents that plan as a flow: the five activities form the backbone, served by the methods of inquiry (Section 3.4) and the methods of design and evaluation (Section 3.5). The dashed return arrow expresses the iteration within the design half: evaluation findings fed back into design during development, while the definitive evaluation ran on a frozen version of the artefact.

The plan reflects three choices: the risk factors (SQ2) and the requirements (SQ3) receive dedicated chapters because they answer different questions (what must be assessed, and how the assessment must behave); the expert interviews feed three chapters (Chapters 4, 5, and 6), because recent industry expertise is essential in this young market; and the artefact is deliberately scoped, with the full conceptual process carried in Appendix E while the thesis implements its foundational block in depth (Chapter 7).

3.4. Methods of Inquiry

The first half of the thesis (Chapters 2 – 6) characterises the problem context and specifies what the artefact must do, using three methods.

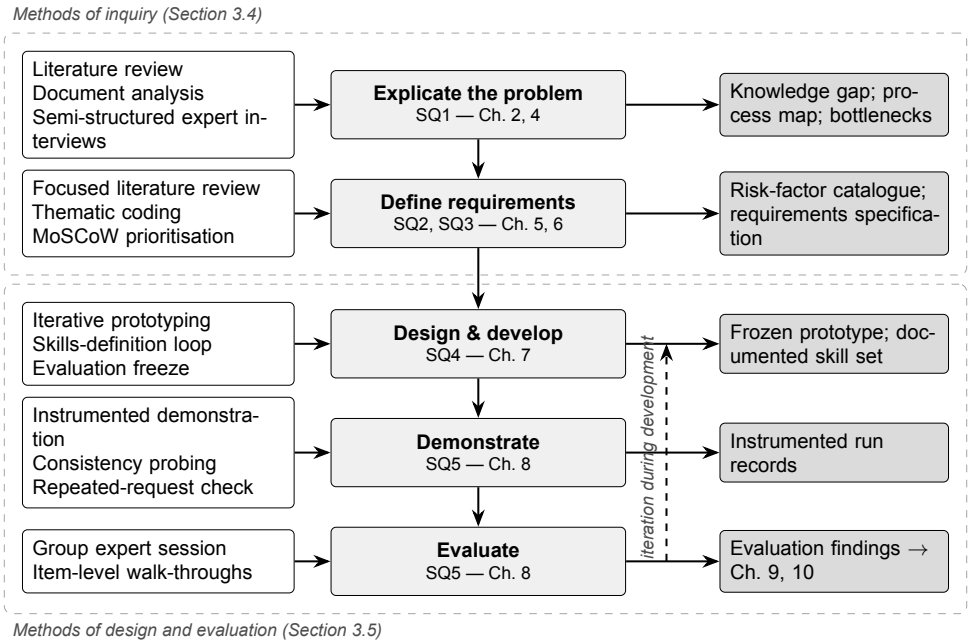

Figure 3.1: Research-flow diagram: the five DSR activities of Johannesson and Perjons [53].

3.4.1. Literature review

Two complementary reviews were conducted: the broad review of Chapter 2, which establishes the conceptual framework, and the focused review of Section 5.1, which extends the risk typologies of Section 2.3 to a traceable, factor-level catalogue. Both were conducted as scoping reviews rather than systematic reviews, because the relevant evidence is dispersed across the BESS, infrastructure-investment, due-diligence, and LLM literatures, and a systematic protocol built for one of those bodies would not reach the others [69]. Peer-reviewed literature was therefore combined with grey literature from identifiable, established institutional sources recent enough to reflect current practice; reference snowballing supplemented database search.

3.4.2. Document analysis

The screening process of Chapter 4 was mapped through document analysis as described by Bowen [12]: an iterative procedure of skimming, reading, and interpreting internal documents to identify themes and process structures. The material comprised internal documents of the BESS developer, accessed under non-disclosure (project pitchbooks, non-binding offer templates, internal due diligence summaries, 4P deliverables, board memos, risk matrices). As a non-reactive source, the documents provided the formal record against which the interview accounts could be triangulated. Informal conversations with the business development, project development, and investment teams validated the resulting process map and surfaced implicit, undocumented decision logic.

3.4.3. Semi-structured expert interviews

Semi-structured expert interviews form the dominant source of primary data, informing Chapters 4 (SQ1), 5 (SQ2), and 6 (SQ3). The record, protocol, and summaries are provided in Appendix B.

Sampling was purposive and gatekeeper-informed: an internal gatekeeper with line-of-sight across the full acquisition workflow (P1) proposed the initial candidates, and the frame covers every functional role active in the acquisition decision, supplemented by external advisors whose cross-client perspective triangulates the internal accounts. Sample adequacy rests on purposive role coverage rather than on demonstrated thematic saturation [83]: the functional roles active in the decision were identified before fieldwork and each is represented. Whether new themes had stopped emerging was monitored during analysis through the stabilisation of the bottleneck and requirement categories, and is reported in Appendix B. With nine participants from one organisation, saturation is claimed for neither the sector nor the full space of screening practices.

Interviews followed a shared opening protocol (role, process insights, risk indicators, requirements) with expertise-specific questioning thereafter. The material was analysed with an approach adapted from the six-phase procedure of Braun and Clarke [13]: familiarisation by listening to each recording in full and capturing the relevant points in structured notes, followed by coding of these notes along four predefined categories (process insights, risk indicators, requirements, other) with inductive sub-codes. The analysis thereby stayed at the level of themes rather than line-by-line transcript coding. Per-interview summaries were returned to each participant for a factual member check. Participants verified the accuracy of what they had said; they did not review the coding, the cross-case themes, the risk catalogue, or the requirements prioritisation, so the check validates the record and not the analysis.

3.5. Methods of Design and Evaluation

The second half of the thesis (Chapters 7 and 8) develops and evaluates the artefact through three methods. In the language of systems engineering, the assessment performed here is validation, not verification: it asks whether the right artefact was built for the practice it serves, not whether it was built correctly against a formal specification [48].

3.5.1. Iterative prototyping with skills definition

The artefact was developed through iterative prototyping, shaped by the choice to implement the workflow with an LLM (Claude Haiku 4.5 and Claude Sonnet 4.6, by Anthropic). Design effort was therefore directed not at model architecture, which is fixed and inaccessible, but at the surrounding software and the structured prompts (*skills*) that each instruct the model to perform one defined analysis task; in the implemented block these are extraction tasks over the checklist items of Chapter 7.

The development followed a skills-definition loop driven by the researcher: a request was sent to the model on real prospect material, the response was inspected against the researcher's knowledge of the case, and where the output was incomplete, mistaken, or poorly structured, the software or the skill was refined and the material rerun. Colleagues at the BESS developer were consulted informally when a response touched knowledge the researcher did not hold; their remarks fed the same loop.

The development history was recorded in a log of the main version changes per development phase, condensed in Appendix D.5; the skill files themselves are versioned, and their frozen texts are reproduced in Appendix D.2. Iteration stopped at the time horizon the MSc thesis allowed, at which point the prototype was frozen for evaluation. Throughout, the LLM was accepted as a black box: only input–output behaviour was observed and manipulated, and the mitigation for the resulting opacity [71] is to document the input side extensively.

3.5.2. Instrumented demonstration and consistency probing

At the freeze, the prompts, checklist version, and model configuration were fixed, and no further design changes were made during evaluation. The frozen prototype was applied to a demonstration set of real prospects from the historical acquisition pipeline, combining teaser-only prospects with full data rooms so that every pipeline phase was observed under realistic input conditions; one data room was held back during development and used for the first time on the frozen version, so that at least one deep run is untouched by design iteration.

The demonstration was instrumented rather than observational: every run records the prompts and documents sent, the per-item results, the detected conflicts, the quote-verification outcomes, the reviewer's actions, and the measured cost and duration. Those records establish what each phase contributes, where extraction fails, and how the human gates are exercised.

Because the demonstration set contained structurally similar prospects, the design additionally permitted consistency probing. All model calls ran at temperature zero, the setting at which the model selects the most probable token and sampling variability is minimised [80]. Identical outputs were not assumed: hosted APIs do not promise them across repeated requests, and repeatability was therefore assessed empirically rather than taken for granted. The evaluation therefore separated two sources of variation. A repeated-request check issued an identical request more than once and measured how much the output changes on its own. Near-identical inputs under the same frozen prompt then allowed item-level comparison of classification behaviour, with the repeated-request check bounding how much

of any divergence is run-to-run variance rather than a difference in the input.

In the terms of Venable, Pries-Heje, and Baskerville [98], the evaluation is ex post and formative: naturalistic in its inputs (real prospects, real documents, the real reviewer role) and artificial in its setting (the prospects were not live decisions at the moment of the runs). It establishes how the artefact behaves under realistic conditions, not predictive accuracy against realised outcomes, which requires the outcome-labelled benchmark identified in Chapter 9 as the natural next study.

3.5.3. Expert evaluation

The instrumented demonstration establishes behaviour; expert evaluation establishes fit. Two formats were used, reported in Chapter 8 and documented in Appendix C.

The first format was a structured group session with practitioners of the BESS developer, combining participants from the interview programme, who could judge whether the artefact addresses the problems they themselves articulated, with newcomers whose fresh eyes tested whether the process is followable. It was organised around three targets that follow from the DSR logic: fit with the screening problem, fit with the requirements, and adoption intent.

The second format was an item-level walk-through, conducted twice: a practitioner with first-hand knowledge of a demonstrated prospect reviewed the filled record item by item, so that extraction quality was checked against personal case knowledge rather than only against the documents. Methodologically this is a concurrent verbalisation over the artefact's output, in the verbal-protocol tradition [28]. The two sessions differed deliberately in depth and domain: one was brief and business-development oriented, the other extended, covering a legal and due-diligence reading with a decline-and-re-run cycle at each gate, so that the artefact's response to expert feedback became observable as well as its initial output. One protocol deviation was accepted: where the expert had not recently worked in the data room concerned, the extraction was shown before the expert's own answer was elicited, so the session audits the record against its citation trail rather than reconstructing it independently, and its agreement figures are upper bounds. In both walk-throughs the researcher was an involved participant rather than a silent observer, a recognised position in interpretive field research provided it is disclosed and mitigated [100]. Verdicts were therefore recorded literally in the prototype's trace, and the positions the experts volunteered are reported separately from those reached jointly. The full session accounts are given in Appendix C.

3.6. Research Quality, Rigour, and Ethical Considerations

The defensibility of this design rests on two things: a build conducted transparently and traceably enough for the contribution to transfer beyond the local context, and empirical work that meets the standards of human-subjects research and confidential-data handling appropriate to a company-embedded MSc thesis.

Research quality and rigour

Four rigour measures operationalise the response to the DSR critique of Section 3.2. *Traceability*: every requirement is linked to the bottleneck and interview evidence that motivate it, every skill to the task and requirements it serves, and every evaluation finding to the corresponding design choice. *Triangulation*: findings were corroborated across data sources where possible (internal interviewees against external advisors, documents against conversations, literature-derived factors against practice-derived indicators), and single-source findings are flagged. *Documented normative selection*: choices that are not algorithmically determined (which factors to carry forward, which requirements to prioritise, which skills to retire) are made explicit with a documented rationale. *Limitations stated in advance*: the interview sample is role-complete but small and concentrated in one organisation, the evaluation is formative and does not measure accuracy or time saving against realised outcomes, and the prototype's outputs are not deterministic; these limitations are revisited in Chapter 9.

Ethical considerations

The research involved human subjects (interviewees and evaluation participants) and confidential organisational data.

Human research consent. The protocol covering the interview programme and the evaluation sessions was approved by the TU Delft Human Research Ethics Committee (application 6960). Informed consent was obtained and archived for every participant, recordings and transcripts were handled on a TU Delft-approved platform, and each participant could withdraw and could review the summary of their own interview before its findings were used.

Confidentiality and data handling. Internal documents were accessed under a non-disclosure understanding consistent with standard practice for company-embedded theses. Prospects are referred to by Greek-letter code names, quantitative details that could identify a transaction are omitted or generalised, and interviewees are referenced by P-code with role descriptions at an aggregation level that does not identify individuals (Appendix B).

Third-party model use and model versioning. The prototype sends confidential DD documentation to a third-party LLM provider (Anthropic) through its API, a data boundary the interview-level measures do not cover. Processing took place under the BESS developer's existing commercial agreement with the provider, under which submitted data is not used for model training by default. That contractual term is one element of a confidentiality assessment and not the whole of it, since retention periods and processing regions are governed separately. The compatibility of the arrangement with the non-disclosure agreements towards selling parties was confirmed within the organisation before any prospect material entered the prototype. As a reproducibility measure, the model versions were fixed for the entire prototyping and evaluation phase, and all reported behaviour is a snapshot of those versions. The governance question of introducing a third-party AI system into a confidential workflow [92] is taken up in Chapter 9, including the alternative of models run on infrastructure under the organisation's own control.

Researcher positionality. The researcher conducted the project as an intern embedded with the BESS developer. The embedding provided access an external researcher would not have, but it also shapes the researcher's view of the process, and gatekeeper-informed sampling may reflect the gatekeeper's view of the relevant population. The same organisation supplied the problem, most participants, the inherited checklist, the demonstration cases, and the intended users, which creates a standing risk of organisational confirmation: what is useful to this company is not thereby valid as research. Usefulness to the organisation is therefore reported as adoption intent, and validity is argued separately from it throughout Chapters 8 and 9. The biases were further mitigated through triangulation, factual member checks, documented normative selection, and the two external advisors (P4 and P9), whose cross-client perspective is independent of any internal account.

4

Current Practice at a BESS Developer: Stakeholders, Process, and Bottlenecks

The literature reviewed in Chapter 2 established the generic development logic of BESS projects, the structural features of investment risk in infrastructure development, and the methods currently used to assess such risk. Together these elements provide a literature-grounded answer to the first sub-question (SQ1). They cannot, however, answer it fully. DSR situates a designed solution within a specific decision environment, and prescriptive design knowledge must be grounded not only in what the field reports but in how the problem manifests in the case under study [53]. This chapter provides that empirical grounding by examining the early-stage investment screening process at the BESS developer introduced in Chapter 1.

One distinction should be drawn before the analysis starts, because it recurs throughout the chapter. The BESS developer realises projects through two channels: it develops greenfield projects itself, and it acquires early-stage projects that prospect developers have originated and offer for sale. This thesis concerns the second channel. In the acquisition channel, the BESS developer acts as the acquiring party and the prospect developer as the selling counterparty, even though both are, in a generic sense, developers of battery storage. The chapter proceeds from environment to process to problem: Section 4.1 maps the actors that shape the screening environment and the tensions between them, Section 4.2 makes the acquisition workflow explicit, and Section 4.3 identifies the recurring bottlenecks from the expert interviews. Section 4.4 translates the findings into the inputs of Chapters 5 and 6.

4.1. Stakeholders in BESS Development

The development stages of Section 2.2 do not unfold in isolation: each phase involves actors whose decisions, information contributions, and institutional mandates determine whether projects advance, stall, or fail [1, 51]. Understanding who these actors are, and how much influence they exert, shows where information asymmetries and coordination failures arise in early-stage screening. To structure the analysis, a power-interest matrix classifies stakeholders by their interest in a project's outcome and their influence over its progression [15]. Drawing on the development literature of Section 2.2, supplemented by the broader infrastructure and investment-risk literature, ten stakeholder groups external to the acquiring party can be identified; the BESS developer itself, as the principal of this research, is treated separately rather than placed in the grid. Table 4.1 summarises each actor's primary interest, and Figure 4.1 maps the groups onto the grid: actors requiring close management combine high power with high interest (the prospect developer, the TSO/DSO, financial institutions), actors to be kept satisfied hold high power with lower project-specific interest (regulators, permitting authorities), and the remaining groups participate stage-dependently or in an advisory capacity.

Table 4.1: Stakeholders in early-stage BESS investment screening.

Stakeholder	Primary interest
Prospect Developer (selling party)	Maximise sale price; minimise post-sale obligations
TSO / DSO	Grid stability and capacity allocation
Financial Institutions & Lenders*	Bankability; covenant compliance
Energy Regulators / NRAs	Market design; tariff regimes
Municipalities & Permitting Authorities	Spatial planning; public acceptance
External Advisors (legal, technical, commercial)	Mandate scope; reputational liability
BRPs & Market Optimisers	Trading margin; portfolio balancing
Local Communities & Public Stakeholders	Quality of living environment
Landowners	Lease income; land-use stability
Battery Technology Suppliers	Order volume; long-term service contracts

*Where the institution is a private equity investor with an equity stake in the acquirer, it holds a dual role, requiring management attention closer to that of a co-principal than of a conventional lender.

Stakeholder Interdependencies and Tensions

Four structural tensions shape early-stage screening in this landscape. The first is information asymmetry: the relationship between the prospect developer and the acquiring party is characterised by information asymmetry: the seller holds the most detailed knowledge of site-specific conditions and has a financial incentive to present them favourably, creating a principal-agent dynamic in which the acquirer must independently verify what it is shown [88, 94]. Second, grid operators combine high influence with limited alignment to project timelines. They hold a legal monopoly on connections, which places the developer in a structural dependency it cannot negotiate its way out of, while connection processing can take years and follows the operator's regulatory mandates rather than commercial schedules [36, 29]. Third, permitting authorities and regulators operate under legal and procedural frameworks that are largely independent of commercial considerations; retroactive policy changes and administrative delays are identified in the literature as critical risk drivers, and their logic resists translation into quantifiable risk indicators [51, 75]. Fourth, participation is stage-dependent in exactly the wrong way: lenders and external DD providers, whose assessments would most directly reduce uncertainty, engage only after substantial development capital is committed, leaving the earliest and most uncertain decisions with the least external verification [74, 88].

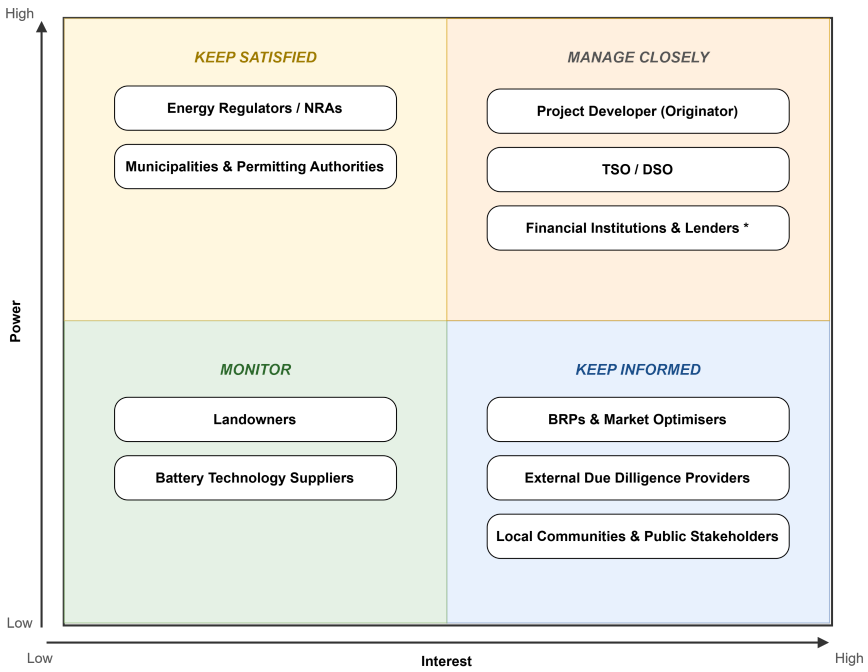

Figure 4.1: Power-interest grid of stakeholders in early-stage BESS investment screening.

Implications for the Prototype Design

Three external actor groups emerge as most consequential to the early-stage screening decision: the system operator, the permitting authority, and the prospect developer. The first two control the grid-connection conditions and the legal preconditions on which a project rests; the third controls both the underlying site conditions and the documentation an acquirer gets to see. These three dependencies correspond directly to three of the four dimensions of the 4P framework that the BESS developer applies in practice (introduced in Section 4.2): *Power*, *Permit*, and *Plot*, with *Profit* as the acquirer's own synthesis of the other three. The infrastructure-investment literature supports the centrality of this alignment: where any of the external dimensions is unresolved, residual idiosyncratic uncertainty remains too high for the asset to be considered bankable [88, 36]. For the prototype this implies that early-screening information requires independent verification or triangulation because it originates with the selling party; that indicators for grid and permit dimensions must capture uncertainty about future actor behaviour, and that the prototype should partially bring forward DD-tasks by structuring and standardising the information that is already available.

4.2. Current Investment Screening Process

The description of the screening process in this section was developed from two complementary sources: a structured *document analysis* of internal materials (project pitchbooks, NBO templates, internal DD summaries, board memos, and 4P deliverables) to reconstruct the formal sequence of acquisition activities, and *informal conversations* with the business development, project development, and investment teams to validate the workflow and surface implicit decision logic (Section 3.4.2). The bottlenecks of Section 4.3 draw on the distinct evidence base of the nine expert interviews.

Two diagrams make the process explicit. Figure 4.2 situates the acquisition workflow within the developer's broader project lifecycle and delineates the scope of this research; it is a high-level overview rather than a formal process model. Figure 4.3 presents the acquisition workflow itself as a multi-actor swimlane diagram, using BPMN-style conventions selectively: actor pools, sequential activities, and gateway decisions in standard form, finer-grained event constructs omitted for readability [18], a selective use that is common in applied process description [79].

Position of Acquisition in the BESS Project Lifecycle

The developer's project lifecycle is organised into seven sequential phases, from initial project identification (Phase 0) to early operations (Phase VI). *Project Acquisition* is the central activity of Phases I and II, spanning the transition from feasibility and viability assessment to the consolidation of the plot, power, permit, and profit dimensions; engineering, contracting, and finance follow after acquisition completes. The case-specific phasing aligns closely with the generic lifecycle identified in the literature (Figure 2.1), adding finer-grained substages within it. This positioning matters for the design: the acquisition decision falls before any engineering or contracting commitment and well before lender DD, yet the assumptions established here cascade through every later phase. As the bottleneck analysis shows, those assumptions are not consistently revisited when the project finance model is built later, which makes the acquisition phase the earliest opportunity for a structured, data-driven instrument.

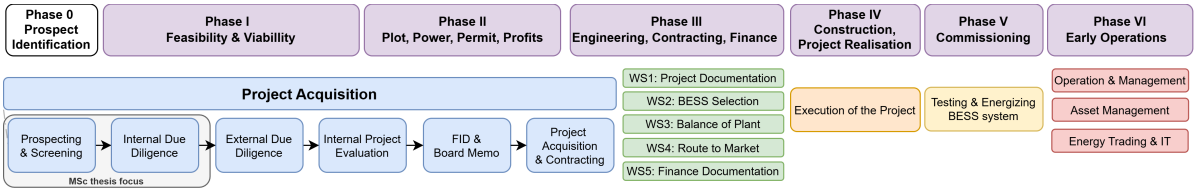

Figure 4.2: The BESS developer's project lifecycle.

The Acquisition Process: Stages and Gates

Within the acquisition workstream, six sequential stages can be distinguished, supported by four explicit decision gates. Figure 4.3 represents this workflow as a multi-actor process model in which five participant groups (the prospect developer, the business development team, the management board, external advisors, and the supervisory board) interact across the stages.

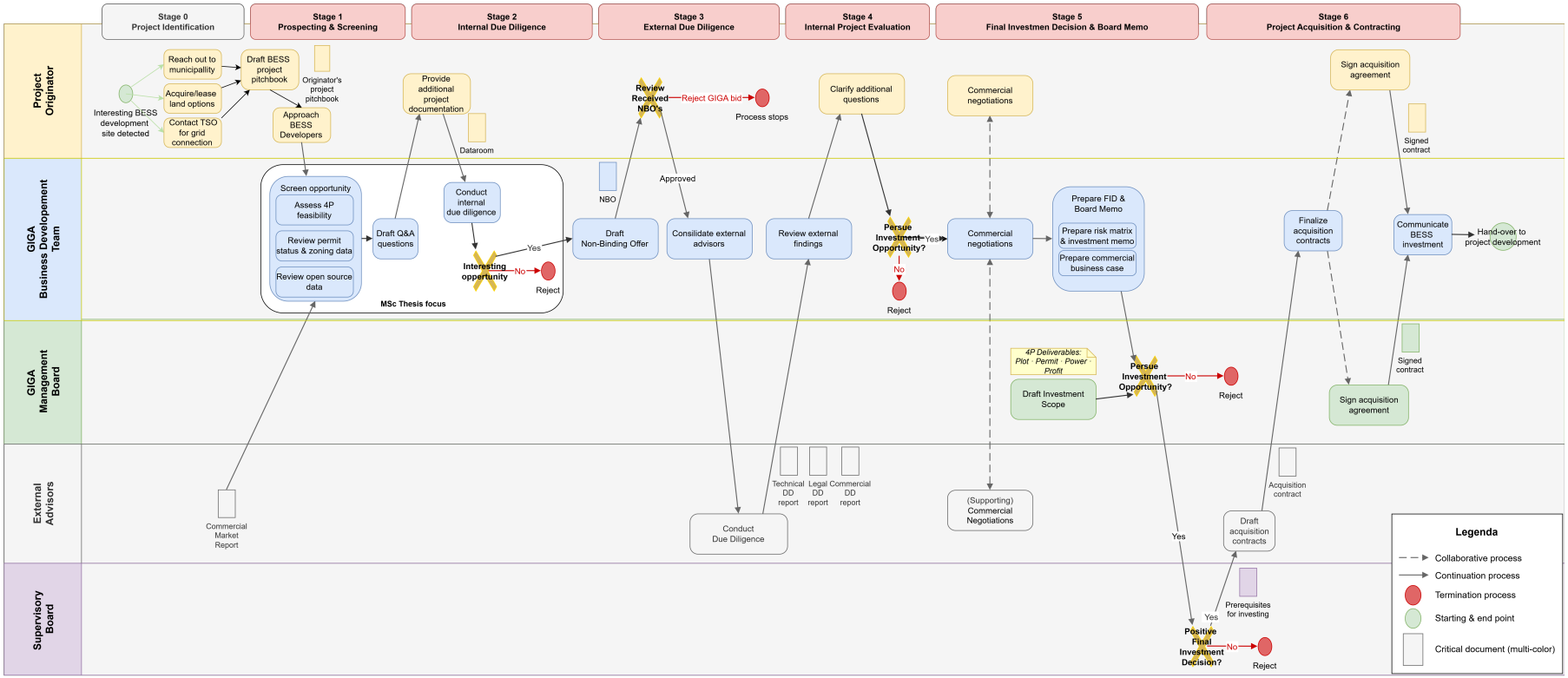

Figure 4.3: Multi-actor process model of the project acquisition workflow.
*The diagram uses BPMN-style swimlane conventions of Chinosi and Trombetta [18] but is not intended as a formally compliant BPMN model.

Stage 0 — Project Identification. The workflow begins with the detection of an interesting BESS development site, typically by the prospect developer, whose preparatory activities include approaching the municipality, securing land options or lease agreements, and contacting the TSO about a grid connection. These activities culminate in a teaser that introduces the opportunity to potential acquirers. The BESS developer's involvement is reactive: opportunities arrive through direct outreach, sell-side advisors, banks, and informal industry contacts.

Stage 1 — Prospecting and Screening. On receipt of a teaser, the business development team conducts a first-pass manual screening organised around the 4P framework: *Plot*, *Power*, *Permit*, and *Profit*. The screening combines internal feasibility assessment, review of permit status and zoning data, and validation against publicly available sources such as TSO network maps. The first gate (*interesting opportunity?*) closes the stage: a negative outcome terminates the process, a positive outcome triggers internal DD.

Stage 2 — Internal DD. The prospect developer provides additional documentation. The team drafts a structured Q&A and conducts a more thorough internal review across the 4P dimensions, resulting in a non-binding offer (NBO). Where the sale is competitive, the NBO is one of several bids and may be rejected, terminating the process externally.

Stage 3 — External DD. If the NBO is accepted, external advisors are mobilised. Technical, legal, and (where commissioned) commercial advisors conduct independent DD on the data room and deliver standalone reports, which the business development team consolidates and reviews.

Stage 4 — Internal Project Evaluation. The consolidated external findings are reviewed alongside the internal results, and clarifying questions are returned to the seller. The second gate (*pursue investment opportunity?*) closes the stage, with the management board deciding whether to advance to the formal commercial negotiations.

Stage 5 — Final Investment Decision and Board Memo. On approval, the team prepares the decision materials (a risk matrix, an investment memo, and a commercial business case) for the supervisory board, which represents the shareholder and issues the final investment decision.

Stage 6 — Project Acquisition and Contracting. Once the FID is positive, the parties finalise the acquisition contracts, drafted by the external advisors. On signature of the share purchase agreement, the project is handed over to the project finance for the route to financial close.

The workflow above is stylised. In practice it is not fully formalised: the company is in a scale-up phase and still developing the procedural infrastructure that would make the process fully repeatable across deals. Sequencing varies with the seller's practices (a data room may accompany the pitchbook or open only after an NDA), and the flow is not strictly linear, since clarification questions can place a prospect on hold until the seller responds. Figure 4.3 should therefore be read as representative of typical practice rather than as a prescriptive specification.

4.3. Five Recurring Bottlenecks in Current Practice

The process map describes the formal structure of the workflow, but not where it systematically produces unsatisfactory outcomes. This section draws on the nine semi-structured expert interviews (internal staff P1–P3, P5–P8; external advisors P4, P9). Five bottlenecks emerge with strong cross-source convergence.

B1. Over-reliance on information supplied by the prospect developer. Opportunities are evaluated almost exclusively on the seller's information, with limited independent verification possibilities and verification capacity. The bottleneck is that the reliance is not made explicit or reduced where it could be. P6 identified it as the root cause of past investment errors, including regulatory risk underestimation on projects in Germany and Belgium, noting that one historical acquisition was evaluated partly on the developer's own analysis, described as "a bit strange" in retrospect. P2 observes that the screening process contains many unverified assumptions, and P9 confirms from a cross-client vantage point that data-room documentation is rarely cross-checked against public sources before contract negotiations begin. This corresponds to the information asymmetry described in the M&A literature [102], which in this workflow remains structurally embedded instead of being progressively reduced.

B2. Silo-based review with no helicopter overview. Internal DD is performed by specialists working in parallel, each on the dimension closest to their expertise. P5 describes the result as structural: each team member reviews a slice without anyone holding the complete view, so important nuances are systematically missed. P6 calls the absence of a helicopter view "the biggest challenge of a growing company", and P9 corroborates from outside: advisors do not read each other's reports, and the financial model is rarely shared in a structured way with the legal or technical streams. The convergence of internal, strategic, and external perspectives makes this the most clearly evidenced bottleneck, and it mirrors the fragmentation that the DD literature identifies as the central quality risk of parallel workstreams [11].

B3. Absence of a formal cross-functional evidence gate. The workflow does have transaction gates, the four decision points of Section 4.2 at which the process may proceed or stop. What it lacks is a documented cross-functional evidence gate: a defined moment at which the technical, legal, and commercial perspectives must converge on one shared record before an investment decision. P3 and P6 identify the gap explicitly, noting that the four-eyes principle is not enforced across functions, with decisions emerging from informal group conversations. P5 links the absence of structured cross-functional review to analytical errors that surface only later.

B4. Assumption inconsistency across phases and advisors. Assumptions are not systematically reconciled with those used later or in parallel DD streams. P5, P8, and P9 describe that revenue curves, CAPEX per MW, OPEX, and interest-rate assumptions that are neither cross-checked nor carried into the project-finance model, creating downstream errors that are expensive to correct. P5 sums the underlying issue up as a lack of "academic rigour" in internal analyses.

B5. Status-focused rather than future-focused evaluation. P6 identifies a final, more conceptual bottleneck: past acquisitions were evaluated against what was in place rather than against what the project must become to be operable and profitable, which he frames as the single most important lesson from past errors. The current process verifies the present state of documentation well but is structurally weaker at forward-looking variables (regulatory trajectories, contract evolution, market structure changes). This is the outside-view discipline that the project-forecasting literature prescribes, and current practice lacks it [32].

The Integrative Screening Problem

Early-stage BESS screening requires synthesising fragmented, partly verifiable, multi-domain information into a defensible go/no-go judgement under time pressure and with limited external assurance. The current workflow distributes the information across specialists, validates only a subset of inputs (Stage 3), and produces its decision artefacts through informal consensus rather than structured cross-functional review. The workflow's capacity to integrate information across actors, dimensions, and time horizons is therefore weaker than the decision context demands. This framing aligns with the literature: infrastructure investment in early stages is characterised by deep idiosyncratic uncertainty, stage-gated information accumulation, and bankability thresholds at financial close [88]. The bottlenecks arise from the absence of an integrative instrument for meeting the demands the literature identifies, rather than from a failure to follow established practice.

4.4. From Bottlenecks to Requirements

This chapter provided the empirical grounding of the design problem: a multi-actor environment with information asymmetry and temporal mismatches concentrated in three external groups (Section 4.1); a six-stage, four-gate acquisition workflow in which early information is dominated by the selling party and external verification arrives late (Section 4.2); and five recurring bottlenecks that motivate an integrative decision support instrument (Section 4.3). Chapters 5 and 6 take these findings forward: Chapter 5 organises the risk factors surfaced by the same interview programme into the catalogue the prototype builds on, and Chapter 6 translates the bottlenecks into the requirements the prototype must satisfy.

5

Investment Risk Factors for BESS Acquisition

This chapter answers research sub-question SQ2, defined in Chapter 3, by constructing and structuring a catalogue of the investment risk factors that an assessment of an early-stage BESS acquisition must cover. The catalogue draws on three streams of evidence: a structured narrative review of academic and grey literature, the nine expert interviews described in Section 3.4.3, and the 4P framework used internally at the BESS developer. The output is a list of 52 risk factors organised into 12 academic categories and mapped onto the 4Ps of Plot, Power, Permit and Profit, following the hierarchy that carries through the thesis: a 4P dimension groups academic categories, a category contains risk factors, and a factor is observed through the checklist items of Chapter 7.

Substantively, the chapter specifies *what* an early-stage BESS acquisition analysis must cover; procedurally, it forms one half of a parallel pair with Chapter 6, which specifies how the assessment must behave. Section 5.1 describes how the catalogue was constructed, Section 5.2 presents the result and its structure, Section 5.3 states the methodological limitations, and Section 5.4 closes with the implications for the prototype.

5.1. Constructing the Risk-Factor Catalogue

The catalogue was constructed in four steps: literature-based identification, interview-based elicitation, merging into a combined candidate catalogue, and a relevance filter that reduced the candidates to a focused working list. Figure 5.1 summarises the steps and their counts; the text below concentrates on the choices behind them.

Literature-based identification

The literature stream was conducted as a structured narrative review with grey-literature supplementation, aimed at breadth rather than exhaustiveness: it identifies the categories of risk relevant to early-stage BESS acquisition with a defensible source for each, rather than performing a systematic review in the PRISMA sense. The starting set consisted of approximately 28 peer-reviewed sources already drawn upon in Chapters 2 and 4 that concern infrastructure investment risk, BESS characteristics, and merger-and-acquisition DD. Snowballing and targeted searches for topics not adequately covered (Netherlands grid congestion, BESS fire safety, public-private-partnership risk allocation, and utility-scale ESS contracting practice) expanded the list to 47. The 19 additions are largely institutional grey literature: international organisations [94, 47, 49, 45, 9], national regulators [29, 10, 91], national research laboratories [19], standards bodies [73, 97], parliamentary research services [43], and one major law firm [8], included only for BESS contracting practice that the peer-reviewed literature does not address.

For each source, risk factors were extracted by close reading. Several sources supplied explicit risk taxonomies that could be transcribed directly [1, 85, 54, 93]; others required interpretive extraction: the

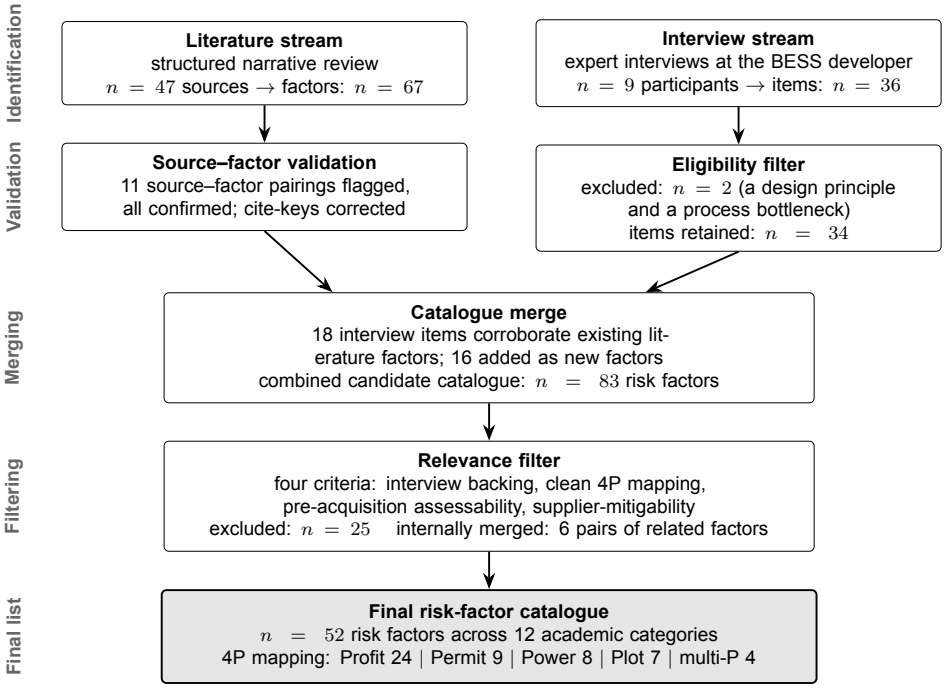

Figure 5.1: Flow of risk-factor identification and selection.

Bayesian Belief Network nodes of Islam et al. [51] and Khan et al. [55] were read as risk factors with causal interdependencies, the due diligence frameworks of Wangerin [102], Bhagwan, Grobbelaar, and Bam [11] and Flyvbjerg [32] were read to extract the risk types they surface narratively, and the risk-allocation matrices of the PPP guidelines [94] and the ESS contracting guide [8] were transcribed factor by factor. Where multiple sources identified the same risk under different names (for example "consenting risk", "regulatory approval risk" and "permit delay"), the factor was consolidated and all supporting sources listed jointly, with sameness judged on the underlying mechanism and impact channel.

The categorisation framework was anchored in the nine-category typology of Abba, Balta-Ozkan, and Hart [1]: Technical, Resource, Policy & Regulatory, Political, Economic & Finance, Market, Social, Grid & Curtailment, and Environmental & Ecological. Three categories were added where risks could not be accommodated without distortion: *Construction*, because the labour, supply-chain and EPC-contractor risks of Shaktawat and Vadhera [85], Taylan et al. [93] and Islam et al. [51] belong specifically to the pre-operational phase; *Legal & Contractual*, because the M&A literature [102, 11] and the contracting guides [94, 8] raise risks that operate at the contract level and conflate awkwardly with regulatory risks; and *Counterparty*, because offtaker credit, manufacturer strength and developer information quality are assessed differently from market analysis. The literature stream yielded 67 risk factors across these 12 categories, each attributable to at least one named source.

Interview-based elicitation

Nine expert interviews were conducted at the BESS developer with participants from acquisition, development, technical, legal and commercial functions (method in Section 3.4.3, record in Appendix B). Their role in this chapter is to elicit the risk factors practitioners actually use when screening prospects, and to corroborate or extend the factors identified from the literature. Each interview-derived factor was classified on a robustness scale that distinguishes saturated themes from individual mentions: *strongly corroborated* (raised independently by all or nearly all eligible interviewees, in convergent terms), *corroborated* (raised by multiple eligible interviewees), *role-specific* (raised by one or two interviewees from their particular vantage point), *singular* (raised by one interviewee but flagged for substantive importance), and *contested* (raised by multiple interviewees with conflicting content). Two items were excluded before the merge because they are not risk factors in the sense of this chapter (one described a design principle for the scoring rubric, the other a process bottleneck); both are taken up in Chapter 7.

The eligibility filter therefore retained 34 items.

Merging into a combined candidate catalogue

The 67 literature-based factors and the 34 interview-based items were merged under two rules: an interview item that described the same underlying risk as a literature factor (the same mechanism and impact channel) was added as a corroborating source on that row, and an item no literature factor covered was added as a new row. Of the 34 retained items, 18 corroborated existing factors and 16 became new factors, giving a combined candidate catalogue of 83 risk factors, each tagged with a preliminary 4P mapping.

Relevance filter and final consolidation

A catalogue of 83 candidates is too coarse to drive an analytical prototype: factors that cannot be evaluated at the screening stage, or that are uninformative in the northwestern European setting, would dilute its output. Four filter criteria were therefore applied: backing from the interview material (directly or through close substantive equivalence), a clean mapping onto one of the four P-dimensions, assessability at the pre-acquisition screening stage, and non-mitigability through the later choice of equipment manufacturer or EPC contractor. The last criterion moves procurement-stage decisions out of the acquisition-stage catalogue: battery-cell-level technical risks (cell degradation, round-trip efficiency shortfall, serial defects) fell entirely on the supplier-mitigable side, while fire and thermal-runaway risk was retained because its consequences for permitting and community acceptance are not removed by a manufacturer choice.

Applying the filter to all 83 candidates excluded twenty-five factors, and six pairs of factors that described variants of the same risk under different names were consolidated by internal merging, leaving the 52 factors of the final catalogue. The interview-coded identifiers were renumbered into the academic-category prefixes, and the numbers of dropped or absorbed factors were not reused, so that gaps in the sequence remain visible as an audit trail. The four-criterion filter was articulated in dialogue with the candidate catalogue rather than designed in advance; the implications are discussed in Section 5.3.

5.2. The Risk-Factor Catalogue

The catalogue resulting from the construction process is presented in Table 5.1. Appendix F provides the same catalogue with additional columns for the industry of the supporting sources and the assessment method recorded where a source provided one. The identifier prefix encodes the academic category: `TEC` = Technical, `RES` = Resource, `POL` = Policy & Regulatory, `PLT` = Political, `ECO` = Economic & Finance, `MKT` = Market, `SOC` = Social, `GRD` = Grid & Curtailment, `ENV` = Environmental & Ecological, `CON` = Construction, `LEG` = Legal & Contractual, `CPT` = Counterparty.

Table 5.1: The risk-factor catalogue of 52 investment risk factors.

ID	4P	Risk factor	Sources
CON-02	Power	Supply chain disruption (equipment delivery)	[93, 8, 94]
CON-06	Plot	Heavy-transport and logistics accessibility	P2, P3, P4, P7
CON-07	Profit	Project schedule slippage and cost overruns	[85, 51, 16, 32]; P1, P2, P4, P5, P6, P8, P9
CPT-01	Profit	Offtaker credit quality (TSO, utility)	[11, 94, 61]
CPT-04	Profit	Prospect developer information quality	[32, 102, 11]; P1, P2, P5, P9
CPT-05	Profit	Counterparty corporate hygiene (KvK, KBO, registries, blacklists)	P1, P2, P5, P9
CPT-06	Plot	Information completeness (’what is missing’)	P1, P2, P4, P5, P6, P9
ECO-01	Profit	High initial capital expenditure (CAPEX)	[1, 81, 19]; P1, P4, P6, P7, P8
ECO-03	Profit	Electricity revenue curve uncertainty	[63, 6, 41]; P2, P3, P4, P5, P6, P8
ECO-06	Profit	Macro-economic exposure (interest rates, lithium and battery prices)	[85, 6, 94, 19]; P2, P6, P8
ECO-07	Profit	Inaccurate economic assumptions (CAPEX, OPEX, inflation)	[85]; P8, P9
ECO-11	Profit	Rising grid connection fees and tariffs	[99, 10]; P1, P2, P3, P5 — P8

continued on next page

Table 5.1 continued from previous page

ID	4P	Risk factor	Sources
ECO-13	Profit	Bankability threshold compliance	P2, P4, P6, P8
ECO-14	Profit	Financial-model assumption sensitivity (discount rate, acquisition-vs-financing mismatch)	[6]; P6, P8, P9
ENV-01	Plot / Permit	Noise impact	[85, 91]; P3, P4, P7
ENV-02	Plot	Flood / hydrology risk	[85]
ENV-03	Plot / Permit	Ecological / biodiversity constraints	[85, 55, 91]
ENV-04	Plot	Climate / weather extremes	[85, 94, 20]; P2, P4, P7
GRD-01	Power	Grid connection unavailability / waiting list	[99, 47, 29, 10]
GRD-02	Power	Non-firm connection curtailment risk	[99, 10]; P1-P9
GRD-03	Power	Grid connection technical constraints (MWh, V)	[85, 37, 89]; P1, P3, P7, P9
GRD-04	Power	Uncertainty in grid operator processing time	[99, 47]; P1, P2, P3, P4, P6, P8, P9
GRD-06	Power	Grid connection contract status (legal certainty)	P1, P2, P3, P4, P5, P6, P8, P9
LEG-01	Plot	Land title and rights defects	[1, 11, 94]
LEG-02	Permit	Incomplete permitting or consents	[11, 91, 74]
LEG-03	Profit	PPA / tolling / capacity agreement terms	[94, 8]
LEG-05	Profit	Information asymmetry vs developer (M&A)	[32, 102, 11]; P1, P2, P4, P5, P6, P9
LEG-06	Profit	Tax structure and transaction risk	[102, 11]; P2, P8, P9
MKT-01	Profit	Data availability/reliability for market behaviour	[1, 94]
MKT-02	Profit	Market design evolution (ancillary service market)	[86, 41, 94]
MKT-03	Profit	Value-stacking constraints (regulatory)	[86, 31]
MKT-04	Profit	Counterparty concentration / offtake dependency	[102, 11, 94]; P8
MKT-05	Profit	Country-level BESS market viability	P1, P2, P3, P6
PLT-01	Profit	Adverse legislative, tax or change-in-law risk	[1, 85, 6, 25, 90, 31, 94]
PLT-02	Profit	Policy retraction (withdrawal of support)	[90]; P2, P3, P5, P6, P7, P8
PLT-06	Profit	Regulatory change as 'information uncertainty gap	P2, P4, P6, P8
POL-01	Permit	Permitting and licensing delay or refusal	[1, 91, 94]
POL-02	Power	Uncertainty in grid extension plans and planning	[1, 29, 47]
POL-03	Permit	Inadequate asset classification for storage	[31, 86, 94]
POL-04	Permit / Plot	Restrictive land-use / zoning rules for BESS	[1, 91]; P1, P2, P3, P4, P5, P6, P9
POL-05	Permit	Permit issuing authority identification	P1, P2, P3, P5, P9
POL-06	Permit	Municipal cooperation signal (principebesluit)	P1, P2, P3, P4, P5, P9
POL-07	Permit	Country and regional permit procedure variation	P1, P2, P3, P4, P5, P9
POL-08	Permit	Municipal inexperience with BESS as new asset class	P2
RES-01	Profit	Revenue resource uncertainty (market price spread volatility)	[63, 41, 6]
RES-02	Profit	Ancillary service price and availability uncertainty	[41, 99, 81, 86]; P1, P2, P7, P8
SOC-01	Permit	Local community opposition / social acceptance	[1, 85, 55, 43]; P1, P3, P7
SOC-03	Permit	Public safety perception (post-fire-incident)	[43, 73]
TEC-04	Power	Thermal management and fire / thermal runaway	[37, 77, 73, 97, 43]
TEC-06	Plot	Site geotechnical / soil conditions	[85, 51, 93]; P3, P4, P7
TEC-09	Plot / Power	Cable route feasibility (length, parcel count, route width)	P1, P3, P7, P9
TEC-10	Plot	Land area suitability for planned capacity	P4, P7

The catalogue contains 52 risk factors across 12 academic categories: the nine of Abba, Balta-Ozkan, and Hart [1] and the three added during the review. Its distribution across the 4P dimensions is markedly Profit-heavy: twenty-four factors map onto Profit, against nine for Permit, eight for Power and seven for Plot, with four factors carrying multi-P labels. This distribution reflects the merchant-revenue model that dominates the Netherlands, Belgium and Germany: the physical asset is technically well understood relative to other infrastructure classes, while the revenue model (wholesale price spreads, ancillary-service stacking, capacity payments where they apply) is genuinely novel and exposed to regulatory and market-structural uncertainty. The concentration in market and institutional categories echoes the infrastructure-risk literature, in which institutional risks are a frequent cause of project failure [68, 1].

The count per dimension, however, understates the informational density of Plot and Permit in a screening context. Many Profit-mapped factors are macro-economic uncertainties (interest-rate exposure, price-curve uncertainty, change-in-law risk) that affect every BESS investment in a jurisdiction similarly; they are best assessed at portfolio level and shape an acquirer's overall stance more than the attractiveness of one site. The seven Plot-mapped factors, by contrast, each decompose into many concrete sub-aspects that can be checked against site documentation and aerial imagery during a screening: elevation, terrain slope, surrounding land use, soil composition, distance to the nearest substation, parcel ownership, heavy-transport accessibility. Several Permit-mapped factors decompose similarly into jurisdiction-specific procedural sub-aspects (the consenting authority, the presence of a *principebesluit* or *Vorbescheid*, the scoping of environmental-impact studies). This site-specificity of early-stage information is what the development literature leads one to expect [36, 88]. In a pre-acquisition screening, Plot and Permit factors therefore yield far more decision-relevant variance across prospects than the counts suggest, which shapes the design of Chapter 7: the analytical scaffold descends into the sub-aspects of Plot and Permit, while most Profit factors are treated as portfolio-level inputs.

Of the 52 factors, the majority are corroborated either by the interview material or by multiple independent literature sources. Factors resting on a single literature source without interview corroboration are visible in Appendix F through their narrow industry tag, so the reader can identify where the evidence base is thinnest.

5.3. Limitations of the Catalogue

The catalogue's epistemic status is bounded by five limitations. First, the review was a structured narrative review, not a systematic one: no pre-registered protocol, no formal inclusion criteria, no documented search strings. The catalogue is traceable (every factor is attributable to a named source) but not reproducible in the way a systematic review would be. Second, the review was conducted by a single reviewer in a single pass, without independent extraction or double-coding; the judgement calls in "same risk under a different name" and in category placement were not externally verified. Third, grey-literature quality was judged by institutional reputation rather than a formal scoring tool such as AACODS. Recency, methodological transparency, and the issuing institution's incentives were weighed informally per source, and the two commercially interested sources were used only for specific factual claims that the peer-reviewed literature does not cover [8, 103]. A formal appraisal of conflicts of interest was not performed. Fourth, the categorisation framework reflects one defensible set of choices; a different reviewer could merge or split categories without changing the substantive factor list. Fifth, the relevance filter was articulated in dialogue with the candidate catalogue rather than designed at the outset; its criteria are stated explicitly, were applied uniformly, and left a preserved audit trail of dropped, absorbed and renumbered identifiers. The catalogue is therefore best read as a structured and traceable starting point for the design of Chapter 7, not as a definitive risk inventory.

5.4. From the Catalogue to a Prototype

The catalogue specifies what an early-stage BESS acquisition analysis must cover; Chapter 6 develops, from the same interview material, the requirements that specify how a supporting prototype must behave. Chapter 7 then positions the catalogue in the layered design: the factors belong to the judgement layer, as the working material of the future assessment stage, while the checklist items that observe them belong to the structuring layer the prototype implements. The observation of Section 5.2 about informational density (Plot and Permit decomposing into tangible sub-aspects, several Profit factors being portfolio-level) shapes how the items are differentiated across the 4Ps.

6

Requirements for Data-Driven Decision Support

This chapter translates the empirical findings of Chapter 4 into a structured requirements specification for the prototype designed in Chapter 7, thereby answering the third sub-question. Section 6.1 describes how the requirements were elicited from the interview programme and the literature. Section 6.2 introduces the IEEE 29148:2018 categorisation used to organise them. Section 6.3 presents the MoSCoW prioritisation and the reasoning behind it. Section 6.4 then presents the 41 requirements per category, and Section 6.5 closes with the priority set that drives the prototype design. The complete specification, including full descriptions and source attribution per requirement, is provided in Appendix G.

6.1. Methodology for Requirements Elicitation

Requirements were extracted primarily from the nine semi-structured expert interviews documented in Appendix B, with additional contributions from the literature reviewed in Chapter 2. The extraction followed the same thematic analysis approach applied to the risk factors in Chapter 5: each interview was coded for explicit and implicit statements of what the prototype should do, how it should behave, what data it should handle, and what it should not do. The resulting items were de-duplicated and consolidated into the present specification.

Cross-interview robustness was assessed using the same eligibility-adjusted denominator as for the risk factors: for each requirement, the denominator is the subset of interviewees positioned by their role and questioning to comment on the topic. This protects against the misleading appearance of low corroboration where a topic was simply outside the scope of a given interview. Robustness levels (strongly corroborated, corroborated, role-specific, singular) carry the same meaning as in Chapter 5.

Two requirement merges were applied during consolidation. The original source-attribution principle (an output-level reliability requirement) and the per-data-point provenance specification (a data-level implementation) were merged into QL-01 because they expressed the same concept at different abstraction levels. The cross-project comparison view and the historical project data integration requirement were merged into FR-05 because the comparison view is functionally dependent on the historical data layer.

6.2. Classification Framework

Requirements are classified using the type taxonomy defined in Section 5.2.8.3 of International Organization for Standardization and International Electrotechnical Commission [48]. This standard was selected because it is the consolidated international reference for requirements engineering, jointly issued by ISO, IEC, and IEEE. It superseded earlier standards including IEEE 830-1998 and IEEE 1233-1998 and is the current authoritative source for the field.

Section 5.2.8.3 of the standard lists the following requirement types as the principal categorisation:

Functional/Performance, Interface, Process Requirements, Quality (Non-Functional), Usability/Quality-in-Use, and Human Factors (in this thesis 41 distributed 20; 5; 5; 7; 4 respectively) . The taxonomy applied in this thesis uses five of these six categories. Process Requirements, defined by the standard as contractual or compliance-driven requirements imposed through the contract or statement of work, is omitted because no requirement in the empirical material strictly fits this definition. The remaining five categories are sufficient to classify all 41 requirements without forcing assignments.

6.3. Prioritisation using MoSCoW

Each requirement is assigned a priority using the MoSCoW method defined by the Agile Business Consortium [3]. The method distinguishes four categories: *Must have* requirements are essential to the deliverable; without them, the project fails. *Should have* requirements are important but not strictly essential; a workaround may exist or implementation may be deferred. *Could have* requirements add value but have lower impact and may be dropped if time runs short. *Won't have this time* requirements are deliberately postponed to protect focus, without prejudice to future inclusion.

The practical tests for each category, drawn directly from the Agile Business Consortium definition, are summarised in Table 6.1, together with the resulting distribution of the 41 requirements across the four priorities.

Table 6.1: MoSCoW prioritisation criteria adapted from Agile Business Consortium [3].

Category	Meaning	Practical test	Count
Must have	Essential requirement. Without this, the project, product, or deliverable fails.	"Can we still deliver a usable or acceptable result without this?" If no, it is a Must.	7
Should have	Important requirement, but not absolutely essential for the current delivery. A workaround may exist or implementation can wait.	"Would it hurt quality or value if missing, but still be acceptable?" If yes, it is a Should.	15
Could have	Nice to have. Adds value, but has lower impact and can be dropped first if time runs out.	"Would this improve the result, but not seriously matter if absent?" If yes, it is a Could.	12
Won't have (this time)	Explicitly out of scope for now. Not necessarily rejected forever, but not included in the current version, sprint, report, or project phase.	"Are we deliberately postponing this to protect focus?" If yes, it is a Won't.	7

Priorities were assigned through researcher judgement informed by three considerations. The first was the empirical robustness of the requirement: strongly corroborated requirements were the default candidates for Must or Should. The second was the value contribution of the requirement to the research question and to the early-stage screening problem the prototype targets; requirements whose absence would defeat the prototype's purpose are Must. The third was the feasibility of implementation within the available research timeline; requirements that are technically attractive but practically difficult at this stage were downgraded to Could or Won't. This combination of robustness, value, and feasibility produces a defensible prioritisation that does not reduce to any single consideration, and it remains a set of documented researcher choices rather than an algorithmic outcome (Section 3.6).

Country and regional configuration support (QL-07) provides a representative example of how feasibility shaped the prioritisation. The requirement is strongly corroborated across multiple interviewees and is conceptually important for any production deployment of the tool. It is nevertheless classified as Could rather than Must because building genuine multi-country configuration capability is beyond the scope of an MSc thesis prototype, and a single-country demonstrator is sufficient to test the underlying design principles.

6.4. Five Requirement Categories

Functional/Performance Requirements

Functional and performance requirements describe what the system must do and how well it must perform when doing it [48]. In the prototype context these requirements cover risk scoring, document processing, public-source automation, regulatory monitoring, validation, and output structuring. Twenty functional requirements were identified, of which two (FR-07 data room ingestion and FR-10 public-source retrieval) are classified as Must; both respond directly to the dominance of seller-supplied information in early screening (B1). Table 6.2 lists all functional requirements with abridged descriptions; full descriptions appear in Appendix G.

Table 6.2: Functional/Performance requirements (abridged). Full descriptions in Appendix G.

ID	Title	Description (abridged)	Priority
FR-01	Per-dimension risk classification with traffic-light output	Green/amber/red classification per 4P and cross-cutting dimension, derived from probability $\times$ impact assessment.	Should
FR-02	Composite project quality scorecard with cost, quality, timing scoring	Score each 4P dimension on three independent axes (cost, quality, timing); aggregate to overall project quality.	Could
FR-03	Phase-relative threshold calibration	Score the same indicator differently depending on project maturity phase: early, mid, or ready-to-build.	Won't
FR-04	Fatal-flaw flagging independent of aggregate score	Flag project-killing conditions separately and override the aggregate score, regardless of other strengths.	Should
FR-05	Cross-project comparison view using historical project data	Display projects side-by-side using the BESS developer's historical project data as benchmark for identifying outliers.	Should
FR-06	Standardised assumption list with validation status	Maintain approximately 30 key model inputs per project, tagged by source type and validation status.	Should
FR-07	Data room document ingestion and structured extraction	Parse data room documents and extract key elements (parties, dates, conditions, restrictions) into structured form.	Must
FR-08	Three-category contract classification (irrevocability, liabilities)	For each contract: classify revocability, extract operational restrictions, flag hidden liabilities.	Won't
FR-09	Absence-of-information detection and flagging	Detect when documentation is absent for a required indicator and flag as a distinct unknown status, separate from a low score.	Should
FR-10	Automated retrieval from public data sources	Programmatic retrieval and structuring of public information per indicator from TSO maps, Kadaster, ENTSO-E, and similar.	Must
FR-11	Cable route geospatial calculation (length $\times$ parcel count)	Calculate shortest cable route, total length, and number of cadastral parcels crossed as quantitative risk proxy.	Should
FR-12	Substation expandability assessment	Assess substation expandability from public TSO investment plans and aerial imagery.	Won't
FR-13	Permit authority identification and policy retrieval	Identify the formal permitting authority and check for additional provincial or waterschap approvals required.	Could
FR-14	Continuous public regulatory document monitoring per active jurisdiction	Monitor public regulatory announcements per active country and translate into business case implications.	Could
FR-15	DD-claim cross-verification against external sources	Cross-verify due diligence claims against independent public sources.	Could
FR-16	Business-case-vs-DD assumption reconciliation	Match financial model line items against findings in legal, technical, and commercial DD reports; flag mismatches.	Could
FR-17	Contextual benchmark via cross-municipality energy-transition activity	Use municipal energy-transition activity (EV chargers, prior renewable permits) as proxy for permit-process speed.	Won't
FR-18	Rejection-reason logging across projects	Record reasons why projects were rejected to preserve institutional knowledge across personnel changes.	Should
FR-19	Internal pipeline transparency reporting	Surface pipeline visibility to the wider organisation, addressing the high silent-rejection rate.	Could
FR-20	Stage-gate cross-functional review support	Structure output as common input document for cross-functional stage-gate review before acquisition decisions.	Should

Interface Requirements

Interface requirements describe how the system is required to interact with external systems or with internal system elements, including human elements [48]. In the prototype context these requirements concern the output formats produced by the system and the connectors through which it ingests external data. Five interface requirements were identified, none classified as Must but two (IF-02 map-based visualisation and IF-03 connector layer) as Should. Table 6.3 summarises them.

Table 6.3: Interface requirements (abridged).

ID	Title	Description	Priority
IF-01	Output format compatible with existing reporting flow	Output integrates with existing DD-report Word and PDF format used for investment proposals.	Could
IF-02	Geographic/map-based primary visualisation	Map-based primary visualisation with overlays (HV lines, substations, cadastral parcels, congestion status).	Should
IF-03	Connector layer to external public sources	Modular connectors to external data sources so additional sources can be added without rebuilding the core.	Should
IF-04	Public-source registry with country-specific mappings	Documented registry of public data sources per country with indicators, update frequency, and access method.	Could
IF-05	Primarily text-based input handling (NLP-orientation)	Prioritise document-oriented processing over numerical modelling, as data room content is primarily text.	Won't

Quality (Non-Functional) Requirements

Quality requirements capture system attributes such as reliability, maintainability, security, and adaptability [48]. In the prototype context the most important Quality requirements concern the reliability of the tool's outputs: source attribution (QL-01) and conservative hallucination handling (QL-04) are both classified as Must. This reliability cluster mirrors, from the practitioner side, exactly the documented LLM limitations that Section 2.5 identified in the literature: hallucination, bias, and prompt sensitivity [65, 104]. Seven quality requirements were identified; Table 6.4 summarises them.

Table 6.4: Quality (Non-Functional) requirements (abridged).

ID	Title	Description	Priority
QL-01	Source attribution and per-data-point provenance for every output	Every output traceable to source; per-data-point provenance metadata required (source, retrieval timestamp, validation status).	Must
QL-02	Reasoning transparency for computed outputs	Expose calculation logic for derived scores so domain experts can validate and override.	Should
QL-03	Up-to-date data is non-negotiable	Data freshness is a core reliability criterion; stale regulatory or market data makes the tool unusable.	Should
QL-04	Conservative AI hallucination handling	Distinguish found, not found, and uncertain; conservative handling of model hallucination.	Must
QL-05	Iterative / modular development with rapid feedback loop	Iterative, modular development with rapid feedback rather than a fully specified end-state design.	Could
QL-06	Confidential / protected processing environment	Operate in a non-public, access-controlled environment given the confidentiality of M&A documentation.	Could
QL-07	Country/regional configuration support	Accommodate per-country logic (NL, BE, DE) within one system rather than maintaining three separate systems.	Could

Usability/Quality-in-Use Requirements

Usability and Quality-in-Use requirements define the basis for the design and evaluation of systems to meet user needs, addressing user performance, satisfaction, accessibility, and efficiency [48]. In the prototype context these requirements concern the readability of outputs for non-specialist audiences, the time saving provided by the tool, and the institutional embedding that makes the tool usable beyond a single owner; the helicopter-view requirement (US-01) is the direct counterpart of the silo bottleneck (B2). Five usability requirements were identified; Table 6.5 summarises them.

Table 6.5: Usability/Quality-in-Use requirements (abridged).

ID	Title	Description	Priority
US-01	Helicopter-view as primary output mode	A 15–20 minute readable project overview suitable for non-specialists, with drill-down available for experts.	Should
US-02	Time saving over current DD process	Compress the current four-week external DD process to a credible same-day initial assessment.	Should
US-03	Institutional embedding (not person-dependent)	Tool must be usable by multiple people and survive personnel changes; not dependent on a single owner.	Should
US-04	Accessibility to non-specialists / multi-disciplinary readability	Output readable by people not deeply familiar with every dimension (e.g. Investment Committee members).	Could
US-05	Long-term revenue forecasting out of scope	No long-term revenue forecasts; the tool flags assumption quality only.	Won't

Human Factors Requirements

Human Factors requirements state the required characteristics for the outcomes of interaction with human users [48]. In the prototype context these requirements articulate the boundary between what the tool does and what humans remain accountable for, and they anticipate, from practice, the accountability and human-oversight characteristics that the NIST framework of Section 2.5 formulates from the governance side [92]. All four address this boundary, with three classified as Must (HU-02, HU-03, HU-04). Table 6.6 summarises them.

Table 6.6: Human Factors requirements (abridged).

ID	Title	Description	Priority
HU-01	Tool must not replace the financial model	Tool does not replace the financial model; it validates inputs to the model.	Won't
HU-02	Tool must not replace external advisor sign-off	External advisors retain independent sign-off responsibility; the tool augments DD, never replaces it.	Must
HU-03	Tool must not make autonomous investment decisions	Investment decisions remain with humans; the tool produces input to decisions, never decisions themselves.	Must
HU-04	Mandatory human review layer for all outputs	All tool outputs pass through human review before influencing decisions.	Must

6.5. Synthesis

The empirical work distilled into 41 requirements spanning five IEEE 29148:2018 categories. The seven Must requirements provide the empirical core of the prototype design discussed in Chapter 7: the prototype must ingest data room documents (FR-07), retrieve information from publicly available sources automatically (FR-10), produce outputs with full source attribution (QL-01), handle uncertainty conservatively without inventing missing information (QL-04), and maintain a clear human-tool boundary in which external advisors retain sign-off (HU-02), investment decisions remain with humans (HU-03), and all outputs pass through a mandatory human review (HU-04). The fifteen Should requirements indicate the next priority layer beyond the minimum viable functionality. The twelve Could and seven Won't requirements bound the prototype's scope for the present research and document explicit choices about what is and is not in scope, providing transparency for future work to build upon.

7

Artefact Design and Development

This chapter answers the fourth sub-question by presenting the artefact: how the risk factors of Chapter 5 and the requirements of Chapter 6 are realised in a working LLM-based prototype. It moves from the conceptual design (a four-block process, of which one block is implemented in depth) to the checklist at the centre of that block, the decision logic of the pipeline, and the implementation with the measures that make it reproducible and auditable. The prompt texts, payload specification, checklist documentation, and development history are collected in Appendix D.2 to Appendix D.5; the process design beyond the implemented block is carried in Appendix E.

7.1. A Four-Block Process for LLM-Integrated Prospect Analysis

The design problem defined in Chapters 4 to 6 is not a single task but a chain: heterogeneous prospect documentation must become structured information, and structured information is the basis for an assessed prospect. Structuring contributes directly to the remedy, since the DD literature treats systematic, structured investigation as the mechanism that reduces information asymmetry and makes assessments comparable across evaluators and prospects [11, 32]. Within that chain, three kinds of work can be distinguished: data retrieval turns documents into facts, and its answers are largely right or wrong; knowledge updating maintains what the organisation knows independently of any one prospect (market conditions, its own assessment preferences) and is refreshed periodically rather than per deal; and judgement weighs structured facts against risk factors and preferences, and is interpretive by nature. Because the three have different correctness criteria, input needs, and cadences, the conceptual design separates them rather than feeding one model call with everything at once. The separation also responds to the documented LLM limitations of Section 2.5: bounded, well-specified retrieval over supplied documents is where language models perform dependably, while hallucination and bias risks concentrate where tasks blur into judgement [65, 104]; keeping the kinds of work apart bounds those risks per step and keeps every output traceable to an improvable process step. The result is the decomposition of LLM-integrated prospect analysis into four blocks that converge on the assessment: filling the checklist (retrieval), creating the market overview and defining the organisational assessment (knowledge updating), and assessing the prospect (judgement), with the risk-factor catalogue of Chapter 5 entering as prepared academic groundwork. Figure 7.1 shows the master process; Table 7.1 summarises the blocks, their cadence, their outputs, and their status in this thesis.

Beyond the decomposition itself, the data room feeds the assessment only through the checklist. An earlier draft routed the data room into the assessment step directly; this was rejected because it makes the origin of a statement in the assessed prospect untraceable and an error impossible to locate in a specific step. Constraining the assessment to structured, human-approved inputs preserves the provenance chain from source document to assessment. The placement of people follows the same logic: every block that touches prospect-specific content puts a human review between the model and the next step (the blue elements in Figure 7.1), while the two memory-building blocks run in the back end and are reviewed at configuration time. This placement operationalises the Must-level human-factors requirements (HU-03, HU-04) at the process level, and the dashed frame in Figure 7.1 marks

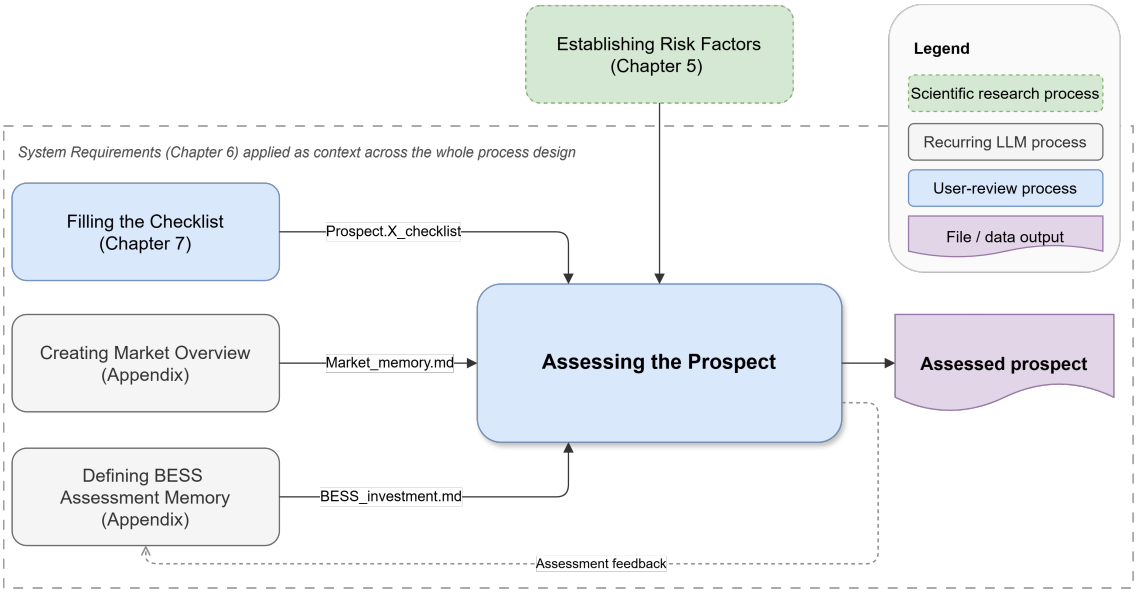

Figure 7.1: Master process of LLM-integrated prospect analysis.

where the remaining requirements of Chapter 6 shaped the processes, from per-output provenance (QL-01) to the preference for explicit absence (QL-04) and periodically refreshed memories (QL-03).

Throughout the design, *skill* is used in the sense of Section 3.5.1 (a documented prompt artefact with a specified task, defined inputs, and an explicit instruction structure) and *prompt* for the runtime object: the assembled instruction text of one specific API request, of which the skill file is the reusable source. Each skill exists as an editable instruction file, reproduced in full in Appendix D.2.

Table 7.1: The four blocks of the conceptual design and their status in this thesis.

Block	Role	Cadence	Output	Status in thesis
Establishing risk factors	Substantive basis of the assessment	Singular	Risk-factor catalogue	Delivered (Chapter 5)
Filling the checklist	Unstructured documentation → structured, attributed record	Per prospect, human-gated	Filled checklist	Implemented as prototype (Ch. 7)
Creating market overview	Public and commissioned market information → market memory	Periodical	Market memory file	Process design (Appendix E)
Defining BESS assessment	Organisational assessment knowledge → assessment memory	Periodical	Assessment memory file	Process design (Appendix E)
Assessing the prospect	Synthesis of the four inputs into a risk assessment	Per prospect	Assessed prospect	Process design (Appendix E)

Why the checklist comes first

Of the four blocks, this thesis implements filling the checklist as a working prototype and carries the others as process designs in Appendix E. The reason is the dependency structure of the chain: the checklist is the foundational block whose output every downstream step consumes, so any inaccuracy that enters here propagates silently into the market comparison, the organisational memory, and ultimately the assessment. Building and hardening the foundation first is therefore a precondition. Two further considerations support this ordering. The structuring task has near-binary correctness and can be assessed item by item, whereas an automated assessment cannot be assessed without first trusting the data layer beneath it. In addition, one block built to a reproducible standard yields more design knowledge than four blocks developed superficially. The practitioner evaluation of Chapter 8 explicitly endorsed this scoping.

7.2. Data Structure and Indicators

The checklist instrument

The data structure at the centre of the implemented block is a checklist of 152 information items that together describe a prospect at the screening stage. The instrument was adopted from practice rather than derived anew: it originates in the manual business-development checklist maintained within the BESS developer, authored by P3, whose items, five-factor structure, and dimensions were taken over roughly as they stood, at about ninety per cent. For this thesis, that internal document was formalised into a standardised schema (version 6): every item received a stable identifier, an explicit extraction definition, a suggested publicly available source, and the process flags that determine how the pipeline treats it, and the researcher assigned each item to the phases in which it is expected to be answerable. What the instrument should contain was thereby treated as given; the consequences of that choice are discussed in Section 9.7. The full schema and descriptive statistics are given in Appendix D.1; Table 7.2 summarises the structure.

Table 7.2: Structure of the checklist instrument (version 6, 152 items).

Factor	Items	Character of the items
Plot	71	Site, land, access, surroundings, cadastral situation
Power	26	Grid connection size, status, route, operator
Permit	22	Zoning, procedures, authority contacts
Selling party	19	Identity, structure, registries, advisers
Profit	14	Fees, tariffs, cost items relevant at screening
11 critical items require explicit human confirmation; 8 items are flagged fully subjective, 9 partially		

The five-factor grouping extends the 4P framework of Chapter 4 with a Selling-party factor, and that extension reflects the bottleneck evidence directly: the prospect developer is the dominant information source in early screening (B1), so the identity, structure, and public record of the selling party are themselves screening information, not background. Each item further carries three phase flags, a `manual_confirmation` flag marking the eleven critical values a human must confirm, and flags marking subjective, location-based, and entity-based items; the flag semantics are specified with the schema in Appendix D.1.

The scoping of the gaps phase is deliberately incomplete: 114 of the 152 items are in scope, and the remaining 38 are excluded because they concern information that exists only in the transaction documents (the deal structure, for instance, or the exact site of a not-yet-public project). A public-source search for such an item cannot return the right answer, but it can return a plausible wrong one (a similarly named company, a nearby project); excluding these items removes a hallucination risk and avoids spending searches where no public ground truth exists.

Relation to the risk-factor catalogue

The checklist and the risk-factor catalogue of Chapter 5 are complementary layers: the catalogue specifies what can go wrong (52 risk factors an assessment must evaluate), the checklist what must be known (the prospect-level facts from which such an evaluation becomes possible). In the four-block design the catalogue enters at the assessment step, while the checklist is the output of the structuring step; many items are the observable indicators of catalogue factors, as Table 7.3 illustrates. Because the prototype implements the structuring step only, the catalogue becomes operational in the assessment stage rather than serving as an evaluation criterion for the prototype itself.

The evidence model

Every checklist item is answered in a fixed evidence model. The LLM assigns exactly one of three evidence states: *found* (the value is clearly stated or shown, with a source), *probably_found* (the documents strongly imply the value; the doubt is explained), or *not_found* (the information is absent; an explicit absence is more useful than a guess). Separately from the state, the LLM records a confidence level, so that a confident absence can be distinguished from an uncertain one. The value itself (*finding*) is kept separate from its derivation and caveats (*reasoning*): a quantitative finding is the bare number with its unit, and any computation or assumption behind it is stated in the reasoning field. Each found

Table 7.3: Illustrative correspondence between risk factors and checklist items.

Risk factor (catalogue)	Example checklist items (indicators)
GRD-06 Grid connection contract status	Power items on connection size, status, and contract documents
CPT-05 Counterparty corporate hygiene	Selling-party items sourced from KvK, KBO-CBE, and Handelsregister
POL-06 Municipal cooperation signal	Permit items on authority contacts and stance (PERMIT-01, PERMIT-02)
POL-04 Restrictive land-use / zoning	Permit item on the applicable zoning plan (PERMIT-03)
TEC-09 Cable route feasibility	Power item on connection length; plot items on route and parcels
ECO-11 Rising grid connection fees	Profit items on grid fees and development contributions (PROFIT-01 to -03)

item carries a structured citation: the exact source filename, the page, and a quote that is later machine-checked against the cited document. This evidence model implements the conservative-handling and provenance requirements (QL-01, QL-02, QL-04) and is enforced identically in every phase through the shared instruction file of Appendix D.2.

7.3. Decision Logic for a Three-Phase Pipeline

The implemented block fills the checklist through a linear, human-gated pipeline of three extraction phases, shown in Figure 7.2 and summarised in Table 7.4. Every phase ends at an approval gate: the reviewer approves the results or declines with written feedback, and a declined phase is re-run with all prior feedback, including per-item rejection reasons, inside its prompt. The pipeline is therefore iterative by construction, but the iteration is always human-initiated and always traceable.

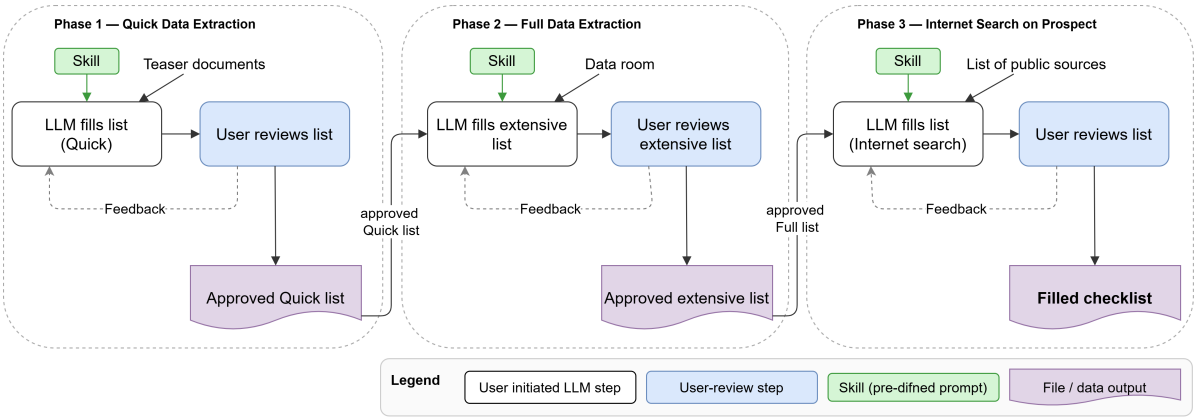

Figure 7.2: The filling-the-checklist pipeline.

The back end adds site detection and adaptive batching. Site detection is a separate call at the start of the quick phase that counts the distinct sites a teaser describes, so that portfolio teasers are split into per-site sub-prospects rather than mixed into one record. Adaptive batching attempts all in-scope items of a phase in a single call and halves the item set only when the response truncates, which keeps a normal phase at one or two calls and bounds the cost of a run (Appendix D.3).

Data-room triage and routing

Real data rooms exceed what a single LLM-API request can carry: rooms of several hundred PDF pages meet hard per-request limits of 100 pages and 32 MB. When a prospect crosses defined thresholds, the engine switches to a three-step mode: an inventory step scans the room locally and estimates per-file page counts; a triage step sends the file listing (paths, sizes, page counts, not the contents) to the cheap classification model, which labels each file's relevance and the factors it informs, with folder structure carrying most of this signal; and a routing step plans each factor's extraction against only that

Table 7.4: The three extraction phases of the implemented pipeline.

Phase	Model	Documents	Items	Purpose
Quick	Haiku 4.5	Teaser only (top-level files)	52	Headline values and site detection
Full	Sonnet 4.6	Data room (routed selection for large rooms)	122	The detailed checklist from the seller's documentation
Gaps	Sonnet 4.6	Routed selection plus live web search for an allow-listed subset	114	Public-source fills and cross-checks of document claims

factor's documents, selected high-relevance-first under per-request caps and a global page budget that acts as the per-run cost ceiling. Every dropped or skipped file is logged by name, and a *not_found* whose factor had dropped documents carries an explicit not-consulted note: an absence under a page budget must never silently imply absence from the data room. Unsupported file formats are likewise skipped and logged; the supported formats, the conversion at upload, and the exact per-call payloads and budgets are specified in Appendix D.3.

Merge logic, conflicts, and ownership

Each phase writes an immutable snapshot of exactly what it returned, and then merges into a single master record per prospect. The merge rules encode the trust hierarchy of the design in three commitments: the human outranks the model, so an edited, confirmed, or ruled-on item is never silently overwritten; later, better-informed phases may challenge but not replace, so disagreements above tolerance (1% for critical items, 5% otherwise) become conflict records that block approval until the reviewer resolves them; and public sources verify documents rather than overrule them. The gates stay honest through two deliberate exceptions: a critical confirmation given before a later phase found new or different data is reopened, and a rejected item that a re-run answers afresh is presented clean, with the rejection reason sent to the model as targeted feedback. Confirmed anchor values (seller name, MW and MWh, voltage level, responsible operator) are injected into the gaps-phase prompt so that public-source lookups target the right company and site; the complete rule set is documented in Appendix D.2 (Table D.4).

Public-source verification

The gaps phase gives the pipeline its independent leg. Items flagged for cross-checking were answered from the data room in the full phase; in the gaps phase they are looked up again from publicly available sources, and the item line in the prompt explicitly carries the prior document value as a cross-check target, with the instruction to report what the public source says on its own merits and never to echo the prior. During the evaluation period, live web search was enabled for an allow-listed subset of these items (seller identity and structure, adviser identity, grid operator, connection status), each in its own bounded call: no documents attached, one search per request, and a hard deadline that cancels the call and records an explicit give-up. The prompt branches honestly on tool availability (a call without the search tool may state consistency with background knowledge but never claim verification), and the merge records a per-item outcome (confirmed, differs, filled, or no public data) that the review interface surfaces as a badge. The mechanism is intended to reduce reliance on claims supplied by the prospect developer (bottleneck B1) by giving them an independent check before they harden into decisions. Chapter 8 reports how far that intention was realised.

7.4. Implementing a Structuring Prospect-Data Prototype

System overview

The prototype is a local web application: a Vue 3 front end for the review workflow, a FastAPI back end that owns the checklist engine (document handling, prompt assembly, API calls, parsing, merging, persistence), and plain JSON files as storage, one folder per prospect. File-based storage keeps every intermediate artefact inspectable and diffable, which serves the traceability measures of Section 3.6. It is a prototype choice and not a production one: access control, encryption at rest, concurrent editing, immutable audit logging, backup, retention, and deletion are not addressed by it, and the production

requirements that would replace it are listed in Appendix D.6. All model inference runs on the Anthropic Claude API, with a small model (Haiku 4.5) for classification work and a larger model (Sonnet 4.6) for extraction, both at temperature zero (Section 3.5.2). Everything else is deterministic local processing: inventory and conversion, routing and budget planning, prompt assembly, parsing, conflict detection, merging, quote verification, and trace writing.

The API is stateless: each call is an independent request, and the model sees only the documents and instructions inside it. All cross-phase knowledge (confirmed anchor values, prior findings to verify, reviewer feedback) therefore exists for the model only because the engine explicitly injects it into a later prompt. This constraint, inconvenient at first sight, is what makes the system auditable: whatever influenced the model literally appears in the recorded request.

Prompt architecture and the answer channel

Every extraction request is assembled from four parts. First, the attached documents (as original files, so that tables, drawings, and figures are read directly). Second, an instruction block composed of a shared general-instructions file followed by the phase-specific skill file with its placeholders substituted. Third, a forced tool schema that defines the exact answer structure per item. Fourth, the generation settings from the skill file's front matter.

The model cannot answer in free text: it must call the answer tool, which makes every response machine-parseable by construction and removes format drift. Three placeholders carry the injected context: the generated item lines, the confirmed anchor values (gaps phase), and the accumulated reviewer feedback. No prompt text lives in code; the skill files are the single editable source, and the exact assembled request of every real call is written to a per-prospect trace together with the measured token usage. Appendix D.2 reproduces the skill files in full and specifies per phase precisely what is sent to the API.

To make the instruction block tangible, the excerpt below shows two fragments of what the model receives in a full-phase request: the role definition from the shared general instructions, and the item block from the full-extraction skill. The horizontal ellipses mark text that sits between the fragments in the real request. The `{{items}}` placeholder is replaced at runtime by one generated line per in-scope checklist item; the line for POWER-03, for example, reads `POWER-03 | Power / Storage capacity (MWh) | extract: capacity | look at: -`. The item lines are therefore where the instrument of Section 7.2 literally becomes the prompt.

```
...
## Who Claude is
You are a due-diligence analyst for Battery Energy Storage System (BESS) project acquisitions
    . Your job in this step is **retrieval only**: you read the uploaded documents and record
     what they say. You never score, rate, or give an opinion on whether the prospect is a
    good deal - a separate step does that later.
...

## Items to fill in this phase
Fill each checklist item listed below - one finding per item, and only for these items. Each
    line is `item_id | factor / topic | extract: <what to extract> | look at: <public source
    >`.

{{items}}
...
```

What the filled record looks like

Table 7.5 shows four items from a filled record, taken from the full-phase run on the fictional prospect Meridian (the same run that Appendix D.4 illustrates from the reviewer's perspective). The four items span the behaviour the design asks for. POWER-03 shows a confirmed critical anchor: a bare value, a verified quote, and the reviewer's confirmation carried from the quick phase. POWER-07 shows the finding/reasoning split on a status question: the finding states the situation, the reasoning explains the boundary (an offer secured, the agreement itself a ready-to-build condition). SELL-11 shows a *probably found*: the documents imply a full exit without stating it, the reasoning reconstructs the inference, and the state hands the verdict to the reviewer. PLOT-30 shows a low-confidence reading of a drawing,

confirmed by the reviewer at the gate. Every row carries its exact source file and page, and every quote was verified mechanically against the document.

Table 7.5: Four items from the filled record of the prospect Meridian.

Item	Finding	Reasoning (abridged)	State, confidence	Source and verified quote
POWER-03 BESS storage capacity (MWh)	800 MWh	—	found, high	Teaser, p. 1: “Standalone 200 MW / 800 MWh” ✓
POWER-07 Preliminary connection agreement signed?	Transport-capacity offer received from TenneT; connection agreement not yet signed	An offer has been secured, but the signed grid connection agreement is listed as a condition for ready-to-build.	found, high	Teaser, p. 2: “A transport-capacity offer for a 380 kV connection to the national grid has been secured” ✓
SELL-11 Seller wants to stay involved?	No post-deal involvement indicated; full exit	100% equity sale via SPA, consideration paid at completion; pre-completion costs reimbursed; no mention of a retained stake or ongoing role.	probably found, medium	Teaser, p. 4: “The Vendor is seeking an investor to acquire 100% of its equity in the Project through an SPA.” ✓
PLOT-30 Plot layout / form	Rectangular grid layout implied by schematic	The schematic shows the site as a rectangular block of containers in a grid; no formal shape or usable-area breakdown is stated.	probably found, low	Teaser, p. 5: “Project Meridian BESS site (~4.6 ha)” ✓

✓ = quote mechanically verified against the cited document.

Verification and traceability layers

The prototype’s output is checkable rather than merely plausible through four mechanisms. Structured citations name the exact source file and page of every found item, powering a click-to-open document view at the cited location. Mechanical quote verification checks every LLM-quoted citation locally against the cited document’s extracted text, and the interface marks each quote as verified or as having no exact match, with unreadable documents honestly left unmarked. The prompt trace records the exact assembled request and measured token usage of every call, doubling as a per-run cost ledger (Appendix D.3). And run instrumentation freezes the start time, duration, and result counts of every attempt of every phase, so that re-runs never erase the record of earlier attempts. Together these layers implement the provenance and transparency requirements (QL-01, QL-02) as properties of the system rather than as habits of its user.

Development and the evaluation freeze

The prototype was developed over twenty logged phases (Appendix D.5), by the researcher-driven refinement loop of Section 3.5.1, repeatedly confronting the design with real prospect material. Several central mechanisms are direct responses to observed failures: adaptive batching replaced a per-factor splitting that re-paid the document cost five times, the forced tool schema replaced free-text output after malformed responses, the honesty branch in the gaps prompt followed a call that claimed a verification it had not performed, and the structural cost bounds on web search followed a server-side search loop that multiplied billed passes far beyond the intended ceiling; the design lessons are distilled in Chapter 9. On 7 July 2026 the prototype was frozen for evaluation (prompts, checklist version, and model configuration fixed), and the demonstration campaign of Chapter 8 ran entirely on this frozen version.

Design constructs and the requirements they serve

Section 3.6 committed this thesis to a traceable chain from requirement to design choice, and Table 7.6 closes that chain for the implemented block: it maps the principal design constructs of this chapter to the requirements they serve and the bottlenecks they address. Whether the constructs satisfy those requirements in use is the subject of Chapter 8.

Table 7.6: Traceability from design constructs to requirements and bottlenecks.

Design construct	Requirements served	Bottleneck addressed
Data-room ingestion, conversion, triage-and-route	FR-07	B2 (shared record replaces parallel reading)
Three evidence states; prefer explicit absence; absence flagged distinctly	QL-04, FR-09	B1 (no invented certainty)
Structured citations; machine-checked quotes; full prompt trace	QL-01, QL-02	B1, B4 (claims are checkable)
Gaps-phase public-source fills and cross-checks	FR-10 (partial: allow-listed live subset)	B1 (independent verification)
Cross-phase conflict detection with tolerance thresholds	supports FR-20	B4 (inconsistencies surfaced, not reconciled silently)
Approval gates; critical confirmations; sticky user ownership	HU-04, HU-03, HU-02	B3 (a structural check per phase)
Decline feedback and per-item rejection reasons into re-runs	FR-18	B2, institutional memory
Retrieval-only role definition; no scoring or opinions	HU-01, HU-03, US-05	scope boundary of the block

8

Demonstration and Evaluation

This chapter answers the fifth sub-question: how the prototype performs under realistic screening conditions, and how domain experts assess its fit with current practice. Following the evaluation design of Section 3.5, the evaluation is formative. It establishes the prototype's behaviour under realistic conditions, its fit with the screening problem of Chapter 4 and the requirements of Chapter 6, and the practitioners' adoption intent. The chapter moves from the evidence base to the instrumented demonstration and the consistency probe, then to the expert evaluation, and closes with the synthesis that answers the sub-question.

8.1. Evaluation Approach and Evidence Base

The checklist instrument was frozen at schema version 6 before the first evaluation activity took place. The prototype software and the skills were refined further between the group session of 1 July and 7 July, when the full configuration (prompts, checklist version, and models) was frozen; these refinements improved the per-item extraction output and left the process that the group session evaluated unchanged. All instrumented evidence in this chapter, and both record walk-throughs, were produced on the frozen configuration of 7 July. Table 8.1 lists the five evidence strands, what each establishes, and what it deliberately cannot establish. Reading them together prevents over-claiming: the quantitative strands describe behaviour, and only the expert strands speak to fit and adoption. An overview of the three practitioner sessions, including their goals, is given in Table C.1 of Appendix C.

Table 8.1: Evidence strands of the evaluation and their scope of validity.

Strand	Basis	Establishes	Does not establish
Instrumented demonstration	Nine pipeline prospects on the frozen prototype	Per-phase yield, conflicts, citation reliability, cost, duration, review activity	Extraction accuracy against ground truth
Consistency probe	Three sibling prospects with near-identical inputs; one repeated identical request	Bounds on classification consistency; nature of divergences	Correctness of individual classifications
Record walk-throughs	Two experts with first-hand knowledge of prospect Beta (P1, 7 July; P2, 27 July)	Content-level audit against expert knowledge; error classes; source-composition evidence	Systematic error rates; independent recall
Group expert session	P1, P3, P5, E1–E3 (1 July)	Fit with problem and requirements; adoption intent; refinement needs	Measured performance in live use
Coverage analysis	Requirements (Appendix G) and bottlenecks (B1–B5) against the design	Which commitments the prototype meets, partially meets, or defers	Sufficiency of the requirements themselves

The demonstration set consists of nine real prospects from the BESS developer's acquisition pipeline, referred to by Greek-letter code names. All nine were processed through the quick phase from their teaser material. Three sibling prospects of one seller group (Alpha, Beta, Gamma), for which complete data rooms were available, were processed through the full pipeline including the gaps phase. The remaining six ran teaser-only, which mirrors the most common real-world situation at first contact. Beta's data room was held back during development and used for the first time on the frozen version, so that at least one deep run is untouched by any design iteration. The two record walk-throughs were designed as complementary readings of the same prospect: P1 audited Beta's campaign record from a business-development perspective, and P2 audited fresh Beta runs from a legal and due-diligence perspective, with the citations of the final run additionally audited by document type.

8.2. Results of the Demonstration

Table 8.2 summarises the campaign. Per prospect it reports, for each executed phase, the number of items classified as found, probably found, and not found; the number of blocking conflicts; the share of citation quotes that the mechanical check verified; and the measured API cost of all runs for that prospect.

Table 8.2: Demonstration campaign on the frozen prototype.

Prospect	Data room	Quick	Full	Gaps*	Blocking confl.	Quotes verified	Cost
Alpha (DE)	yes	20/0/32	29/40/53	10/51/29	0	69%	$1.68
Beta (DE)	yes	18/0/34	69/18/35	13/28/20	1	59%	$4.51
Gamma (DE)	yes	15/3/34	63/19/40	22/40/18	0	54%	$3.87
Delta (DE)	–	12/4/36	–	–	0	31%	$0.06
Epsilon (DE)	–	12/0/40	–	–	0	75%	$0.06
Zeta (NL)	–	37/0/15	–	–	0	81%	$0.20
Eta (DE)	–	18/0/34	–	–	0	88%	$0.06
Theta (DE)	–	17/12/23	–	–	0	71%	$0.08
Iota (DE)	–	15/2/35	–	–	0	62%	$0.06

* The quick and full columns sum to their phase scopes (52 and 122 items). The gaps-phase counts report the items that the phase attempted at run time, a subset of its 114-item scope; the campaign logs do not preserve the per-prospect composition of the unattempted remainder.

What each phase adds: A teaser typically yields 12–20 found items of the 52 quick-scope items; one unusually information-rich teaser reached 37 (Zeta). Full data-room extraction lifts the found count to 63–69 of 122 full-scope items on two of the three deep-run prospects. The difference between these figures indicates what the pipeline contributes: most of the information a screening record requires is absent from the teaser and must be extracted from the data room. That this holds without any task-specific training data corroborates, in a new setting, the central practical finding of the construction-risk literature on LLM extraction [27].

Output variance is explainable: The third deep run (Alpha) found only 29 items with 40 probables, less than half of its siblings' yield. The instrumentation explains why: seventeen files in that room were skipped as unsupported formats in this run. The extraction therefore ran without a document class that carries substantial weight in due diligence. Two properties of the design are relevant. Yield is bounded by format coverage, which is a concrete, fixable limitation. And the system surfaces the gap rather than hiding it: the skipped files are logged by name, and affected absences carry a not-consulted note, so a low yield reads as an incomplete input rather than as a confident empty room.

The independent leg produces signal: Across the three deep runs, the public-source cross-check returned 7 confirmations, 7 material differences, 6 fills of items the documents left open, and 1 case without public data. Thirteen of the 21 checked items (62%) were thus corrected or extended relative to the seller's documentation. This is the quantitative counterpart of bottleneck B1, and of the information asymmetry that the due-diligence literature places at the centre of acquisition screening [102]: information supplied by the prospect developer, checked independently, does not always survive the check.

Cross-phase consistency and conflicts: The deep runs produced 18–24 recorded differences between phases per prospect, essentially all informational; a single blocking conflict occurred across the whole campaign. Teaser and data room therefore largely agree, and where they do not, the comparison machinery of Section 7.3 caught the disagreement instead of silently overwriting one side (bottleneck B4 at the data layer).

Citation reliability: The mechanical quote check verified 59% of quotes literally across the three deep runs (for example 46 verified against 32 unmatched on Beta) and up to 88% on clean, digitally authored teasers; the lowest share (31%, Delta) belongs to heavily tabular material with hardly any full sentences. The unverified share is not a hallucination count: it includes legitimate quotes from documents whose text layer resists exact matching. Its value is diagnostic. The reviewer sees, per item, whether the cited evidence could be machine-confirmed, and the aggregate share indicates how far a given document set can be trusted on citations alone.

Human review activity: The reviewer in this campaign was the researcher rather than an independent domain expert, so these counts describe how the gates were exercised and not how effective they are in expert hands. On the deep runs the reviewer made 15–17 manual edits per prospect, confirmed the critical values (9–10 accepted per prospect, with items marked unavailable where the stage did not yet allow confirmation), and ruled on probable findings predominantly as unsure (37, 18, and 17 items respectively), reserving confirmation for items within the reviewer's own knowledge. Given that the reviewer was the researcher rather than a domain expert, the high share of unsure verdicts is the methodologically honest pattern: the interface records the boundary of the reviewer's knowledge instead of laundering it into confidence.

Cost and time envelope: A quick run costs about $0.06–0.20 and completes in 25–56 seconds. A full deep pipeline costs $1.68–4.51 per prospect, with the longest single phase at 16 minutes. The entire campaign measured $10.58 in API cost. The economic interpretation is developed in Chapter 9; the demonstration-level point is that API inference cost is not the binding constraint on adoption.

8.3. Consistency Under Near-Identical Inputs

The demonstration set contained a natural probe. The three sibling prospects were quick-extracted from near-identical inputs: one PDF (containing the selling party information) was similar across all three, and a spreadsheet with an identical structure whose row values differ per project. All calls ran at temperature zero with the same frozen prompt, which minimises sampling variability as an explanation for divergence (Section 3.5.2). What remains is input difference, context sensitivity, and residual run-to-run variance. The repeated-request check sized the last of these: the identical quick request, issued twice on Zeta in separate runs, produced wording variation across several findings and shifted two of the 52 items (3.8%) between *found* and *probably found*.

Of the 52 quick-scope items, 45 (87%) were classified identically across all three sibling prospects. The seven divergences fall into two categories. Roughly three are genuine data differences and represent the design working as intended: the connection-length item, for instance, differs because the spreadsheet rows genuinely differ. Roughly four are inconsistent classifications of near-identical evidence, of which the repeated-request check suggests about half is within the range that repetition alone produces. Table 8.3 shows the three clearest cases.

Table 8.3: Divergent classifications of near-identical evidence in the consistency probe.

Item	Evidence	Alpha	Beta	Gamma
POWER-03 capacity	Sheet states "TBD" in all three	found ("TBD")	found ("TBD")	probably found
PLOT-02 town	"Undisclosed" (Beta, Gamma)	found	not found	probably found
SELL-18 adviser	The shared PDF names the adviser	found	found	not found

The probe supports two conclusions. First, borderline judgements flip even when sampling variabil-

ity is minimised: across repetition of the identical request (3.8% of items) and across near-identical neighbouring context, as the miss on identical source material shows (SELL-18). Second, a substantial share of the observed inconsistency is a specification gap rather than model noise. The evidence model of Section 7.2 has no rule for explicit non-values: a document that literally states “TBD” can defensibly be read as a found fact (the seller states the value is undecided, which is itself relevant) or as not found (no value exists). Because both readings are defensible, the model oscillates between them. A one-sentence placeholder rule in the shared instructions closes the gap; it was identified during the campaign and deliberately not applied to the frozen prompt, since patching mid-campaign would have broken comparability across runs.

The probe also bounds the interpretation of Section 8.2. Recording an explicit “TBD” as found inflates found counts, so the yield metrics mix extracted values with extracted absences. As a consequence, cross-prospect outcome comparisons require item-level inspection before they can support conclusions. Finally, early-phase variance propagates: quick findings become the anchors that steer the gaps-phase lookups, so a small upstream inconsistency compounds downstream. The methodological status of near-duplicate probing beyond this study is discussed in Chapter 9.

8.4. Expert Evaluation

The prototype was evaluated with practitioners in two formats: a structured group session with the BESS developer’s project development team (P1, P3, P5, and newcomers E1–E3), and two record walk-throughs of prospect Beta (P1 and P2). The group session evaluated the process; the walk-throughs audited the content the prototype produces. The conduct of all three sessions, including dates, durations, participants, and goals, is documented in Appendix C (Table C.1). Table 8.4 condenses the principal findings; the subsections below give the evidence behind each.

Table 8.4: Principal findings of the practitioner evaluation.

Finding	Source	Relation to the design
Structuring unstructured deal documentation is the precondition for any later assessment; participants expect duplicated parallel reading to fall away	Group (P1, P3)	Confirms the scope decision (Section 7.1); addresses B2
Three-state evidence model and per-item verification judged “very useful”	Group (P1)	Corroborates FR-09, QL-01, QL-04
Review obligation should be tiered by item criticality; some items are subjective	Group (P5, E1); P1 walk-through	Refines HU-04; extends the evidence model
Public-source step should be restricted to a whitelist of trusted sources	Group (P1, P3)	Matches the built allow-list (Section 7.3)
One extraction error: seller motivation attributed to the wrong party (PLOT-17)	Walk-through P1	Entity-confusion class; caught at the HU-04 gate
Extraction frequently right in substance but cited to seller-composed summaries instead of underlying evidence (61% of quotations)	Walk-through P2	B1 recurring inside the data room; QL-01 incomplete as specified
Checklist items are the prompt; under-specified items produce under-specified extraction (11 items identified)	Walk-through P2	Instrument re-specification is the first development priority
Feedback re-runs act on corrections, not on general sourcing principles (3/17 moved)	Walk-through P2	Source preference must be governed in the document set, not the prompt
Adoption endorsed, gated on accuracy and time saved against a known baseline	Group (P5, P3)	Defines the follow-up study (Chapter 9)

Group session

The participants recognised the conversion of unstructured deal documentation into a standardised record as addressing a real constraint in current practice. P1 called the checklist-filling step “a sort of vital step …to first make from unstructured data, structured data”. P3 articulated the change in the division of labour: where several team members currently read the same documents in parallel, “the machine can now do that, and one of these persons still has to …read everything and validate the

information so it's readily available for the others." The gain is the removal of duplicated reading, not of human reading, which maps onto bottleneck B2 and matches the hybrid division of labour reported in the construction-risk literature, in which the model supplies breadth and the human supplies depth [66]. A second recognised benefit is institutional memory: logging findings and rejection reasons creates an overview the organisation currently lacks, and it holds even if the interpretive parts of the pipeline are never automated (FR-18).

On the requirements, the session corroborated rather than discovered. The three-way evidence distinction and the per-item verification were judged "very useful" (P1), confirming the reliability requirements the interviews had already marked as non-negotiable (FR-09, QL-01, QL-04). P5 restated the boundary of trust drawn in the interviews and refined it: "I think AI can extract anything. What I would do is …even if AI believes it's 100% certain, to still have a human check", proposing that the review obligation be tiered by item criticality rather than applied uniformly (a refinement of HU-04). E1 added that some items are inherently subjective and need a representation beyond the three-state model; eight items in the schema are already flagged as such. The participants further converged on restricting the public-source step to a whitelist of trusted sources, which the built allow-list mechanism of Section 7.3 implements; P3 supplied the list of sources practitioners use most.

Walk-throughs of the filled Beta record

Business-development reading. P1 audited Beta's campaign record against first-hand project knowledge. For the large majority of reviewed items, the extracted values and classifications matched P1's knowledge, and P1 considered the record a usable screening basis. One substantive extraction error surfaced: on the item recording the seller's motivation to sell (PLOT-17), the model attributed a motivation that belonged to a different party in the transaction chain (the landowner rather than the selling entity). The error belongs to the entity-confusion class already observed during development, occurred on an item flagged as subjective and entity-based, and was caught and corrected at the review gate. P1, who had also reviewed the three sibling runs of Section 8.3, remarked on being positively surprised by their consistency.

Legal and due-diligence reading. P2, who was closely involved in the due diligence when the BESS developer acquired Beta, audited fresh runs of the frozen prototype on that prospect, including a decline-and-re-run cycle at each gate (Appendix C.3). The dominant finding was that the extraction was frequently correct in substance while citing the wrong class of document. Across the full phase the model preferred a single seller-authored timeline overview to the underlying documents from which that overview had itself been compiled. P2's objection is a due-diligence argument: "Due diligence is there to validate whether what the seller says is correct …Whether it is true follows from the underlying documents." A seller's statement is not worthless, only insufficient; the overviews record the existence of a study rather than its content: "It does not say what there is; it says that there is something."

A manual audit of the final run's citations quantifies the objection, and the run's instrumentation locates its cause in the document funnel (Table 8.5). Of all quotations, 61.3% came from documents the selling party had composed, with one timeline overview supplying over a third (Table C.3); P2's indicative target ranges for an adequate record are 40 to 50% document coverage and 5 to 10% seller-composed reliance. These are one practitioner's volunteered heuristics rather than an external standard, and what counts as adequate depends on document redundancy and materiality. The coverage gap is partly architectural (the routing budget caps what reaches the model, and the triage step tagged every candidate the same), and partly the model's preference among the documents it did receive. The clearest instance concerned a critical value: the record carried a superseded headline capacity from the seller's overview, a derived price-per-megawatt figure had been computed against it, and the authoritative letters from the transmission system operator were present in the room but never reached the model.

This finding extends bottleneck B1. The independent public-source leg addresses over-reliance on developer-supplied information from outside the data room, and Section 8.2 shows it producing signal. The walk-through shows the same asymmetry recurring inside the data room, between the seller's summaries and the evidence they summarise; taking the outside view on developer-supplied claims is precisely what the due-diligence literature prescribes [32]. The design does not currently distinguish

Table 8.5: Document funnel of the audited prospect Beta walk-through run.

Stage	Documents	Share of room
In the data room	232	100%
In supported formats (after file-type filter)	217	94%
Routed to the model (after triage and page budget)	44	19%
Cited in the record	11	4.7%

these document classes, so its provenance requirement (QL-01) is satisfied at the level of naming a source without being satisfied at the level of naming the right one, a gap the citation-generation literature does not surface because it treats provenance as an output property [34].

The second development point concerns the instrument. P2 named the specification of the checklist items as the largest remaining development need: “If you regard the left-hand column as a prompt, then this is a bad prompt.” In this design that is literally true, since each in-scope item is compiled into the extraction prompt. Eleven under-specified items were identified and reformulated in the session (Table C.5), including permit items named after Dutch statutory instruments that cannot be found on a German prospect, where a not-found verdict carries no information. The decline-and-re-run cycle sharpened the picture: of seventeen rejected items, the three whose rejection reasons contained a concrete factual correction moved to a better source, and the fourteen carrying only the general instruction to prefer non-seller sources did not. Feedback transmits specific corrections more effectively than general principles; source preference must therefore be governed in the composition of the document set. Further error classes (unsupported citations, findings contradicting their own reasoning, precision lost between phases because confirmed values are not injected into the full-phase prompt) are catalogued in Table C.6, with their causes and fix directions consolidated in Appendix E.

P2’s positive findings were specific: the model retrieves the most obvious relevant material even on unclear questions (“it still gets to warm”); roughly three quarters of the probable findings were estimated correct, off the cuff and scoped to the items the model itself flagged as uncertain; the model performs comparatively well where it declares its own doubt; and the per-item citation with click-through was the working mechanism of the review, as it had been for P1. The session’s verdict: the process is on the right track, and the instrument is not yet thorough enough for the record to be relied on in DD.

Adoption verdict

The practitioners judged the checklist-filling process a favourable change relative to current practice, subject to one clearly stated evidential gap. P5: automating the filling of checklists “makes a lot of sense”; asked directly, P5 confirmed it would “make a lot of sense to start using it”, subject to the refinements proposed in the sessions, and P3 concurred that the effort should focus “100%” on filling the checklist, endorsing the scope decision of Section 7.1. The requested refinements were concrete: iterative re-runs as new documents arrive (P1), a question-and-answer side process that drafts questions to the seller from identified gaps (P3), item prioritisation by influence (P1), and criticality-tiered verification (P5); all are consolidated in the development roadmap of Appendix E.

The evidential gap was named with equal precision. P5: “What I didn’t get yet is what is the quality of the checklists that come out and what is the time that we are going to save by doing this.” Both quantities are what a genuine adoption decision turns on, and P5 proposed the corresponding test: run the tool on completed projects, so that accuracy and time saved can be judged against a known baseline. This defines the follow-up study of Chapter 9. A governance concern from E3 (does the tool replace over-reliance on advisors with over-reliance on the tool?) was answered by P5 and is supported structurally by the design: probable findings must be ruled on and critical values confirmed, so the reviewer must still consult the underlying material. The tool compresses the manual effort of pre-analysis; it does not remove human accountability from it, which is the governance boundary the NIST framework requires organisations to draw explicitly [92].

8.5. Synthesis: Coverage and Boundaries

Most requirements are met by construction rather than by evaluation: the design chapter already maps each design construct to the requirements it serves (Table 7.6). What the evaluation adds is the cases where use confirmed, qualified, or corrected the picture. Table 8.6 reports exactly those.

Table 8.6: Requirements where the evaluation added evidence beyond the design.

ID	Status	Evidence	What the evaluation established
FR-07	S	Demonstration (§8.2)	Real teasers and rooms converted at negligible cost; yield bounded by format coverage
FR-10	P	Demonstration (§8.2)	*Must* met partially, as a recorded deviation: the allow-listed live subset produces signal (62% corrected or extended) and is extensible; per-source connectors deferred
QL-01	S	Walk-through P2 (§8.4)	Satisfied as specified (file, page, checked quote per finding); incomplete as intended, since a citation may name the seller's own summary or a document that does not contain the quote
QL-04	S	Probe (§8.3)	Conservative handling holds; one boundary gap (explicit non-values: "TBD") identified and left unpatched in the frozen version
HU-04	S	Both walk-throughs	Review gates enabled correction of the errors the participating experts detected; two error classes are invisible to model confidence entirely, and undetected errors cannot be excluded
FR-09	S	Demonstration	Not-found state & not-consulted notes behaved accordingly

Against the bottlenecks of Chapter 4: B1 is addressed by provenance and the independent cross-check while the P2 walk-through shows the same asymmetry recurring inside the data room, unaddressed. B2 is addressed at the data layer, as P3's division-of-labour reframing confirms; the helicopter view above that layer remains future work. B3 is partially addressed: the pipeline installs gates per phase, while the cross-functional decision gate belongs to the unimplemented assessment stage. B4 is addressed within the pipeline through anchors and conflict detection with explicit tolerances, not yet for quantitative market assumptions. B5 is not addressed it is a property of the assessment layer.

The answer to the fifth sub-question is therefore two-sided. On the dimension the prototype was built to improve, the observed behaviour and the expert judgements point in the same direction:

- the process becomes more structured for the user, and so does the information available to them;
- real teasers and data rooms become standardised records at negligible marginal API cost ($10.58 for the whole campaign, excluding reviewer time, engineering, and production controls);
- the independent check corrected or extended 62% of the claims it verified;
- four of the five documented bottlenecks are addressed at the data layer;
- the practitioners endorse both its value and its scope, and asked to start using it.

On the dimension due diligence ultimately requires of a record, it falls demonstrably short, and the evaluation locates why:

- 61% of audited quotations rested on seller-composed documents against a practitioner benchmark of 5 to 10%, with 4.7% of the room cited (25% of the routed documents);
- classification of borderline evidence remains sensitive to specification, context, and repetition itself (8% divergence across near-identical siblings; 3.8% across identical repeated runs);
- extraction yield is bounded by format coverage and the routing budget;
- the two quantities a real adoption decision turns on, output quality and time saved against a known baseline, remain unmeasured.

The observed shortfalls are consistent with deficiencies in the specification of the instrument and in the composition of the documents placed before the model, both of which are properties of the design. The evaluation varied neither model nor routing, so it does not isolate these effects from model capability. They are addressable within the design, and none of them was established as addressed by this evaluation. These delimitations set the agenda for Chapter 9.

9

Discussion

Chapter 8 reports what was observed; this chapter argues what it means. It opens with the four positions this thesis takes on its own evidence, each with its boundary and beginning with the one the prototype exposes most sharply, then reflects on what the research design enabled and excluded, distils the contribution into design propositions, draws the implications for a BESS developer, takes position on processing counterparty-confidential material through a public LLM, and closes by bounding the claims and opening the research directions.

9.1. Four Positions on the Findings

Negative evidence requires coverage assurance

The sharpest problem this research surfaced concerns what an absence means. The most consequential asymmetry the prototype exposes is not between right and wrong answers but between kinds of claims. A *found* is a claim about a passage, verifiable in seconds through its citation; a *not found* is a claim about a search process, asserting that a document set, read through a particular funnel of formats, triage, and budgets, contains no answer. Its credibility therefore depends on coverage (which formats were readable, which documents were routed, how complete the room was), none of which is visible in the answer itself. The design's not-consulted notes are a first step; the funnel of Table 8.5 shows how much they must carry. The evidence model does not yet operationalise this: a reviewer sees a bare *not found* and not the funnel behind it. The position connects the evidence model, the routing architecture, and the trust problem the interviews called the classic problem of due diligence (P1, P2): if users come to trust not-found verdicts while coverage is silently bounded, the tool's most valuable output becomes its most dangerous one. Negative findings should therefore be reported with their coverage, not merely with their state; this is where the evidence model most clearly needs its next state.

Specification as the binding constraint

In the failure modes observed here, several errors were traceable to item specification and to document routing. Neither model nor routing was varied experimentally, so this is a within-case interpretation rather than a measured ranking of causes. Where output fell short, the cause traced to something the designer controls: an item that did not say what counts as an answer, a document that never reached the request, a confirmed value not injected into the next phase. The measured variance points the same way: with sampling randomness removed, 3.8% of items still flipped state between identical runs and 8% across near-identical siblings, and half of that divergence traces to items with no rule for the situation they met. P2's summary of the largest development need was the left-hand column of the checklist, not the model behind it.

The position is bounded in two ways. Model capability was not isolated as an experimental variable: one frozen model pair was evaluated, so the findings do not establish that specification would remain the binding constraint across other models, document types, or the judgement tasks this thesis delib-

erately did not automate. And sharper specification is not free: a highly specified instrument encodes its authors' assumptions rigidly, can suppress weak signals that fall outside the schema, and can go stale as markets move. The defensible form of the position is therefore that specification is where an organisation's tacit screening judgement becomes explicit, reproducible, and improvable, including its possible bias, and that in this research it was the control the designer could most readily exercise over what the model produced. Whether it dominates model capability is a proposition for the comparative study proposed in Section 9.8. Integrating an LLM therefore consists largely of specifying precisely what is to be extracted. That specification work is domain-bound and organisation-specific, it accumulates over time, and it is not performed by adopting a more capable model.

Provenance as an end-to-end property

The engineering literature makes models verifiable by requiring citations and checking them [34, 59], and this thesis adopted that framing in its provenance requirement (QL-01). The evaluation shows output-level citation checks to be necessary but insufficient: every audited finding named a real file, a page, and a checkable quote, and the record still failed the purpose provenance serves in due diligence, with 61% of quotations resting on documents the selling party had itself composed, because the model preferred the seller's summaries and the routing had never placed the underlying instruments before it. Instructing the model to prefer better sources changed little; a preference cannot be exercised over absent documents, and long-context capacity does not dissolve the problem, since evidence placement itself affects whether a model uses it [60].

In document sets assembled by an interested party, provenance must therefore be treated as an end-to-end property of the pipeline: source admissibility, document selection, retrieval, extraction, and citation verification stand or fall together, and the earliest stages dominate. Requirements of the QL-01 kind must specify the class of admissible source, and the mechanism that composes the document set (here, a triage step and a page budget built as a cost ceiling) must be recognised as the component that actually decides what the record can cite. This reframes bottleneck B1: over-reliance on the prospect developer's account recurs inside the data room, between the seller's summaries and the evidence they summarise, which is precisely the distinction due diligence exists to interrogate [102, 32]. A tool that reduces the asymmetry between seller and acquirer can, at the same time, create a new asymmetry between those who understand what the routing admitted and those who only read the generated record; citation transparency does not reveal what was excluded before generation.

The constraining split between data-structuring and judging

The practitioners endorsed concentrating the machine on structuring and keeping judgement human, and the errors landed where that split predicts: on items close to interpretation, and where the model reasoned around its instructions instead of retrieving. Yet structuring is not judgement-free: deciding what counts as evidence, whose role an entity plays, when an implication is strong enough for *probably found*, and when an absence is meaningful are all interpretive acts. The contribution of the design is not the removal of judgement from the structuring layer but its constraint, visibility, and reallocation: the evidence states force the interpretive verdicts into the open, the confidence field separates them from the facts, and the gates hand them to a human. This thesis therefore holds that an LLM can carry the structuring layer of screening under supervision today, and that the judgement layer should not be automated until the structuring layer beneath it is demonstrably reliable, because judgement built on an unaudited record inherits every unnoticed defect of that record.

9.2. Reflection on the Research Design

The three-cycle view of Hevner [39], introduced in Section 3.2, frames the audit: Figure 9.1 instantiates the cycles for this research. The relevance cycle ran in both directions and more than once: the problem and requirements came from the field, and the prototype went back to it three times, the second walk-through returning not a refinement but a requirement the prototype satisfies as specified and fails as intended, complete with the expert's own diagnosis and a discriminating experiment; feedback of that kind is why evaluation belongs inside a design cycle rather than after it. The rigour cycle drew on Chapter 2 and returns the design principles, two evaluation methods (near-duplicate probing and the repeated-request check), a fully documented prompt set answering the opacity critique of design-science prototypes [71], and the provenance position above. The design cycle is documented in the

development log: twenty phases closed by a freeze, Beta's data room held back entirely, and a known improvement deliberately left unpatched to preserve comparability.

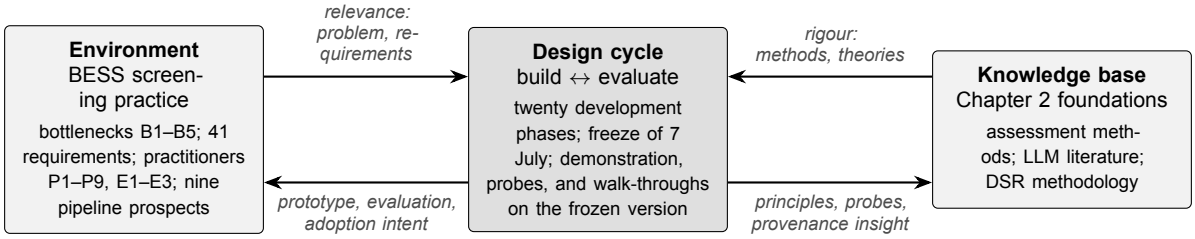

Figure 9.1: The three cycles of Hevner [39], instantiated for this research.

An audit of the three cycles does not by itself make the reflection critical. The more informative question is what this research design enabled and what it excluded. Committing the main question to an LLM enabled deep design work on one technology and excluded a technology-neutral answer to whether an LLM is the best instrument for the task. Organisational embedding gave access to tacit knowledge and real material that interviews alone would not have surfaced, while making the inherited checklist and workflow appear more natural than alternatives. Iterative prototyping exposed mechanisms (routing, provenance, anchor wiring) no interviewing would have revealed, freezing strengthened comparability at the price of preserving known defects, and adoption intent demonstrates relevance rather than effectiveness. The principal trade-off was therefore depth against comparison: the mechanisms of one design became observable, while commitment to that design limits what can be claimed about alternatives.

9.3. Contribution as Prescriptive Knowledge

The confrontation with the knowledge gap of Chapter 2 is the task of the conclusion; what belongs here is the form in which the contribution transfers. Following Iivari [44], the local build is distilled into prescriptive knowledge: the reframing that the four contextual conditions of early-stage BESS screening need not be met by one method but distribute over a layered process, and eight design principles for integrating an LLM into confidential, high-stakes document work. The principles fall into three families (epistemic integrity, human authority and state continuity, and inspectability with enforceable constraints), and each is anchored in an observed failure or a measured behaviour rather than in intention.

Epistemic integrity.

1. **Return an explicit absence rather than an unsupported value, and verify every quotation mechanically against its cited document.** An absence that is stated can be investigated; an invented value cannot be distinguished from a correct one without re-reading the source.
2. **The system should not report a check it did not perform, cite a source it did not read, or discard work without recording it.** Each of these was observed as a failure during development before the corresponding rule was added to the instructions.
3. **Evidence should be ranked by its independence from the party it evidences.** Where the document set is assembled by an interested party, the distinction that carries information is between a claim corroborated by that party and a claim corroborated independently of it.
4. **Where a specification is compiled into the prompt, the precision of the specification bounds the precision of the output.** Instruments inherited from human practice carry ambiguity that human readers resolve tacitly and that a machine reader converts into unrecorded latitude.

Human authority and state continuity.

5. **Human-confirmed values should persist and be injected into subsequent requests.** Once a person edits, confirms, or rejects an item, no later machine output should overwrite it without surfacing the disagreement, and because a stateless API knows only what is supplied, confirmed decisions have to be carried forward deliberately.

6. **Statelessness should be treated as a design property rather than a constraint.** Because each call starts from nothing, everything the model knows is supplied on purpose and therefore appears in the recorded request.

Inspectability and enforceable constraints.

7. **Everything the model receives should be inspectable after the fact.** Instructions live in editable files rather than in code, every request is logged, and nothing reaches the model that the record does not show.
8. **Cost limits should be enforced programmatically through page, call, and time budgets.** A limit expressed only in the instructions can be exceeded by the model's own behaviour, as the web-search incident of Appendix D.5 showed.

9.4. Practical Implications

For the BESS developer, the path from proof of concept to adoption is ordered, and the order follows from the findings:

- **Re-specify the checklist items first.** The items are the prompt; their ambiguity is the model's latitude. This is the highest-leverage work and it requires domain experts, not engineers.
- **Type the data-room documents by author.** Seller-composed summaries must be distinguishable from the instruments they summarise before the record can carry due-diligence weight.
- **Carry confirmed values into every extraction phase,** so that a correction made at one gate constrains everything after it.
- **Then measure against a known baseline.** Accuracy and time saved on completed prospects are the quantities an adoption decision turns on; the practitioners defined this test themselves.

The full development roadmap, including the prototype-level refinements of Table E.1, is consolidated in Appendix E. Adopting the tool would change the organisation, not only the tooling. The operating model the participants envisage is one in which the machine performs the first pass and one person validates a single shared record, where several people now read the same documents in parallel. This is an adoption hypothesis rather than an observed result: no live workflow change, reviewer time, or rework was measured. That reallocation is an explicit workflow decision, and it asks something of the people as much as of the process: specialists must grow accustomed to reviewing and challenging a machine-produced record rather than building their own view from scratch, and to treating feedback to the tool as part of the work itself. The compounding effect is institutional memory: the log of what was extracted, corrected, and declined survives staff changes, answering a concern the interviews raised repeatedly. A caution belongs next to that benefit: a standardised instrument privileges the risks already in its schema, so a mature deployment needs an explicit channel for observations that fit no item, or the checklist will slowly narrow the organisation's attention to yesterday's project archetype.

Beyond the case, the evaluation suggests guidance for any organisation weighing LLM integration, stated here as a proposition rather than a measured law: integration succeeds where the goal of each step is defined sharply enough to check, and the design and development effort an application needs grows with the extraction accuracy its output must deliver. Figure 9.2 sketches the proposition for the applications encountered in this research; the dashed line is the proposed relationship, drawn conceptually rather than estimated from data. A design above the line demands more accuracy than its checks and balances can secure, and hallucinations and mistakes pass into the output; a design below it wraps a tolerant task in more control than it needs, so reviews multiply, verification fatigue and rubber-stamping set in, and the added constraints can even interfere with the model's own useful judgement. Enriching a database with rough estimates tolerates error, so light review suffices; filling a screening checklist informs capital decisions, so it warrants evidence states, verified citations, and human gates. Supporting due diligence end-to-end is, on the evidence of this research, achievable rather than speculative, although far beyond the effort scale of an MSc prototype: it demands tight control over instructions and document sets, and verification designed in from the start. The findings of this thesis lean towards this relationship; establishing it empirically is future research (Section 9.8).

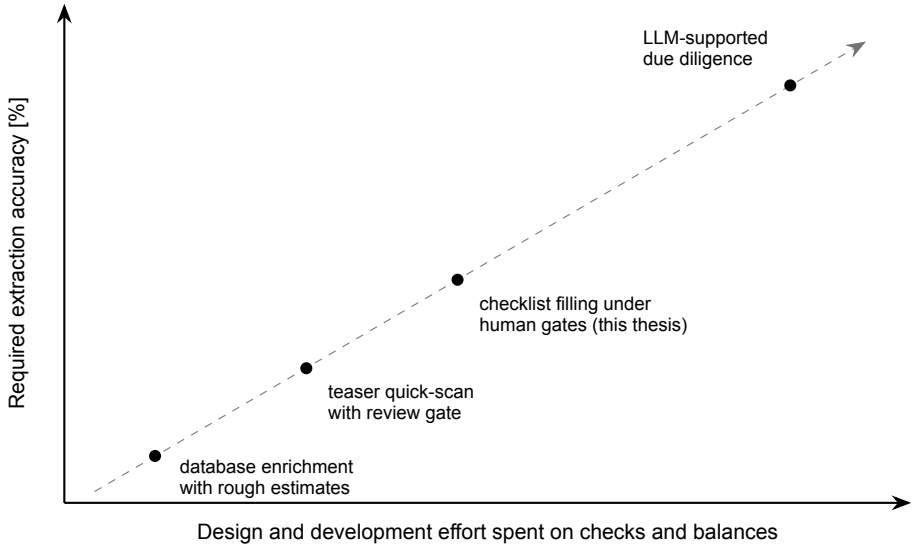

Figure 9.2: Positioning LLM applications in screening and development effort.

9.5. Public API or In-House Model?

This research processed confidential acquisition material through a public LLM API, under the BESS developer's commercial agreement with the provider and with data excluded from model training (Section 3.6). That was a defensible choice for a prototype; whether it remains the right choice for organisation-wide use is a governance question in its own right [92].

The sharpest risk is not the developer's own data. During screening, a data room is counterparty-confidential information whose permitted processing is governed by a non-disclosure agreement: it belongs to a transaction that has not happened, and it does not become the acquirer's until the deal closes. Sending it to a third-party model provider therefore creates a disclosure channel for information the organisation does not own, and the terms of each NDA determine whether that channel is permissible at all. In this research the compatibility was confirmed beforehand; a rollout cannot treat that as a one-off check, because every new seller relationship brings its own terms, and acting on them per counterparty is part of the process design, not a legal afterthought.

The agreement also does not by itself make the workflow secure. The security review with the BESS developer's specialists (6 July 2026) defined the hardening production would add: defences against prompt injection through seller-controlled documents, the leading risk class for LLM applications [76] that the forced answer format constrains only on the output side; input sanitisation; structured PDF processing; zero-trust handling of untrusted documents; and observability of the traffic around the harness. None of these depends on where the model runs, and moving the model in-house resolves none of them by itself. The limitation is real and current: seller-controlled documents are adversarial inputs by construction, and the prototype was not tested against prompt injection, hidden instructions, malformed files, or cross-prospect leakage before confidential material was processed through it. On that ground the present design should be called a research prototype and not a responsible production deployment. A red-team test set for that purpose is specified in Appendix D.6.

Where the model runs is, moreover, a continuum rather than a binary choice. A shared public API, zero-retention or hardened API access, a private endpoint, a controlled cloud tenancy and a fully self-hosted model each strike a different balance of capability, vendor dependence, audit rights, retention, cost and operational burden. What changes along the continuum should steer the choice: self-hosting improves confidentiality control and version stability, but not what the model knows or how it behaves, so the epistemic controls of this design remain necessary at every point. On the evidence of this research, the extraction task needed tight specification more than frontier capability, which weakens the capability argument for staying fully public, while the NDA exposure strengthens the control argument. The position this thesis takes is pragmatic. At prototype and early-adoption scale, public API use under a data-processing agreement is proportionate, provided NDA compatibility is verified per counterparty.

The same findings yield a hypothesis for the production phase: once the process becomes load-bearing for acquisition decisions, a public API may no longer satisfy the confidentiality and control demands, and the organisation should then re-evaluate its place on the deployment continuum. That re-evaluation should also revisit requirement QL-06, which asked for confidential handling of prospect data and was deliberately prioritised as a *Could* for this proof of concept.

9.6. Generalisability

The DSR strategy is that of Iivari [44]: solve one concrete problem well, then distil knowledge others can reuse. What transfers must be differentiated, because the four layers of the contribution travel differently. The architecture (a layered, phase-gated, provenance-first process around an exchangeable item list) may transfer to any acquisition context organised around a data room, in other M&A industries as much as in energy. The design principles may transfer wherever an LLM is put to confidential document work. The checklist and its source hierarchy do not transfer automatically: they encode one asset class, three jurisdictions, and one organisation's practice, and the prototype was deliberately built so that a different list can be swapped in, which makes re-specification the price of every transfer. The measured performance figures do not transfer at all without replication: they belong to one frozen model pair, one seller group's data rooms, and a small demonstration set. Transfer is most plausible where the work is document-centred, recurring information needs can be defined, authoritative evidence exists, and decisions pass through identifiable gates; the implementing organisation must specify precisely what it wants extracted, because the domain knowledge is carried by the instrument rather than by the surrounding software.

9.7. Limitations

The limitations are organised by the validity claim they touch.

Construct validity: The instrument was adopted, not derived: about ninety per cent of the items, factors, and dimensions were inherited from the internal business-development checklist, with the thesis adding schema, scoping, and flags (Section 7.2). Whether the checklist represents the information a BESS developer actually needs from a prospect developer's documents is a question this thesis did not answer, and because the items are compiled into every prompt, the limitation propagates into every measured result; a better-thought-through checklist would in all likelihood improve extraction immediately. Related, the meaning of *not found* is bounded by coverage (the coverage-assurance position of Section 9.1): the yield metrics mix extracted values with extracted absences, so they are not accuracy figures.

Internal validity: Several mechanisms could explain individual observations besides the diagnosed cause: routing and format coverage bound what the model could cite (44 of 232 documents routed, tagging the same relevance); the local document filtering was engineered as a cost ceiling and never investigated as a variable; and the researcher participated in both walk-throughs [100], with verdicts recorded literally and the dominant findings volunteered by the experts, but without independence. The conclusion that the model responds to specific corrections rather than general principles rests on one feedback round on one prospect, with routing as an alternative explanation.

External validity: One organisation, one screening culture, three countries, one seller group behind all three deep runs, and one frozen pair of models. The demonstration set is small because a live acquisition pipeline is small; conclusions about how good the model's findings are rest on that base and must be read accordingly.

Reliability: Outputs are not deterministic: identical requests shifted 3.8% of items between evidence states, so single runs carry irreducible variance. The manual citation audit and the tool's automatic citation log return different totals for the same run; the discrepancy is unresolved, and the two counts serve different purposes. All reported behaviour is a snapshot of the frozen model versions, and hallucination is heavily suppressed by the evidence states, the never-invent rules, and the quote checking, but not eliminated: an early gaps-phase version once claimed a verification it had not performed, which is what led to the honesty rule now in the prompt.

Evaluation validity: The evaluation is formative: there is no ground-truth benchmark; the walk-throughs are output-first audits whose agreement figures are upper bounds, and P2's three-quarters esti-

mate is an off-the-cuff impression scoped to the items the model itself flagged; adoption intent is not adoption; and only one process configuration was built, so no claim is made about alternative designs. Evaluating one's own work with colleagues carries obvious risks, reduced but not removed by the mixed audience, the recordings, and the explicit call for criticism.

Governance and ethics: NDA compatibility was confirmed once, within one organisation, and does not generalise per counterparty (Section 9.5). Security against adversarial documents was reviewed but not tested. The human gate protects human decisions without validating them: a mistaken manual edit is as durable as a correct one. Held against the NIST trustworthiness characteristics [92], the prototype is strongest on accountability, transparency, and explainability, partially established on validity and reliability, and untested on security, resilience, and fairness across jurisdictions and document languages. The prototype-level refinement backlog behind these limitations is consolidated in Table E.1 of Appendix E.

9.8. Future Research

The engineering continuation of the prototype is consolidated as a development roadmap in Appendix E; the directions below are the academic agenda this thesis opens, each grown out of the findings and limitations above.

Decision-centred derivation and outcome-linked validation of the screening information model: The instrument was adopted from practice; future research should derive it from the decisions it serves, working backwards from go and no-go gates to the information that materially changes them, and then validate the model longitudinally against realised outcomes (land control, permits, grid milestones, business-case preservation), including rejected prospects to expose false negatives. Such work must keep four validities apart: correct extraction, a defensible assessment on the information of the moment, outcome prediction, and decision rationality under the uncertainty of the time; a correctly screened project can still fail later, and longitudinal research should not read every failure as a screening error.

End-to-end provenance and the assurance of negative findings: The provenance position of Section 9.1 defines an experimental programme: vary source admissibility, document typing, ranking rules, and routing budgets, and measure how findings and source composition respond, beginning with the ablation of seller-composed summaries. A not-found is a coverage claim, so evidence models that report coverage alongside state (including an explicit not-applicable) should be evaluated against the three-state model used here.

Comparative architectures and assurance intensity: This thesis shows that an LLM can perform the structuring task, not that it is the best instrument for every part of it. A comparative benchmark of general-purpose LLMs, smaller specialised models, layout-aware document models, rule-based extraction, and hybrids on one acquisition document set, varying specification detail, feedback specificity, and review intensity, would test the effort-accuracy proposition of Figure 9.2 directly. Metrics should be risk-weighted rather than aggregate: precision and recall per criticality class, the false-not-found rate, citation support, consistency across runs, cost, and reviewer workload, including the fatigue effects that additional review cycles may introduce.

Human–machine complementarity and organisational learning: How cross-disciplinary judgement, silo behaviour, accountability, and expertise retention shift when specialists stop reading in parallel and start reviewing a shared machine-maintained record is an open organisational question. The organisational memory this design proposes needs equal scrutiny: records that preserve expertise can preserve outdated assumptions and bias with the same fidelity, so how captured knowledge is contested, versioned, and retired is part of the research question.

Epistemic control at the boundary between structuring and judgement: The assessment stage will require the model to combine transaction evidence, market knowledge, and organisational preferences, and the development log already shows parametric knowledge leaking into extraction. How reliance on each source can be controlled, exposed, and made to abstain, and whether the judgement layer is better served by hybrids of LLM-structured evidence with the interdependency-capable methods of Chapter 2 (Bayesian networks, multi-criteria methods) than by an LLM alone, is the question on which any extension beyond structuring depends.

10

Conclusion

This thesis set out to answer how an LLM-based screening instrument can be designed to convert unstructured prospect documentation into a human-reviewed screening record that supports early-stage investment risk assessment for utility-scale BESS acquisition. The answer was built in five steps, one per sub-question, and this chapter assembles them. It states the positions this thesis takes on the evidence of Chapters 4 to 8, confronts the knowledge gap of Chapter 2 with what has been delivered, and closes with the societal contribution and the link to the CoSEM programme.

Answer to the Research Question

SQ1: How does early-stage investment risk assessment for BESS project acquisition currently work, and what are its recurring bottlenecks? There is no uniform, established way to assess early-stage BESS investment risk. Practice rests on structured expert judgement: at the BESS developer, a six-stage acquisition workflow with four transaction gates, in which specialists read the prospect developer's documentation in parallel and converge informally on a go/no-go without a shared cross-functional evidence gate. Four conditions explain why more modern, data-driven methods have not taken over. Utility-scale BESS investment is a young niche with a short and heterogeneous transaction history, the available information is predominantly qualitative, the risk dimensions are interdependent, and the screening window is measured in weeks. None of the five method families reviewed in Section 2.4 satisfies these four conditions at once. Critical inputs depend, moreover, on external actors (above all the TSO and the municipal permitting authorities) whose own planning uncertainty prevents reliable estimates at this stage, which the interviews show to be inherent to early-stage development rather than a defect of one organisation. The tractable part of the problem sits in the data: screening information arrives fragmented and largely qualitative, and processing it manually makes the process slow and hard to standardise. Nine expert interviews documented five recurring bottlenecks (B1–B5): over-reliance on seller-supplied information, silo-based review, no formal cross-functional gate, assumption inconsistency across phases, and status-focused evaluation. Together they define an integrative screening problem: fragmented, partly verifiable, multi-domain information must be synthesised into a defensible judgement under time pressure.

SQ2: Which investment risk factors are relevant for early-stage BESS acquisition, and how can they be structured for screening use? Because the asset class is young, no factor-level risk framework for BESS acquisition existed in the literature; category-level typologies from adjacent infrastructure and renewable-energy domains were the closest available starting point [1, 68]. This thesis therefore constructed one. A structured review and practitioner elicitation produced 83 candidates, which screening-stage filtering reduced to a catalogue of 52 investment risk factors across twelve categories, from inaccurate economic assumptions (ECO-07) through climate extremes (ENV-04) to uncertainty in the TSO's grid-extension planning (POL-02). Uncertainty converges on a BESS prospect from many directions at once, which is why a screening must hold all of it in view. The factors are organised in a hierarchy that carries through the design: a 4P dimension (Plot, Power, Permit, Profit) groups aca-

demic risk categories; a category contains risk factors; and a factor is observed through checklist items. Structurally, the risk mass is Profit-heavy in count (24 of 52 factors), but Plot and Permit factors carry most of the prospect-specific, checkable variance at the screening stage, which is where a screening instrument earns its keep. The catalogue itself belongs to the judgement layer of the design: it specifies what an assessment must eventually weigh, while the checklist items that observe its factors belong to the structuring layer that this thesis implemented. The catalogue is thereby both the substantive answer to this sub-question and the specification against which the future assessment stage must be built.

SQ3: What requirements should a data-driven decision support prototype satisfy? The empirical message of the requirements elicitation is consistent: support for data-driven investment screening should make the process more efficient, standardised and rigorous, keep the overview of a prospect up to date, and convert unstructured deal documentation into structured data, while the industry expert stays in control. Elicitation across the nine interviews produced 41 requirements in five IEEE 29148 categories. Seven form the design core: ingest the data room, retrieve from publicly available sources automatically, attribute every output to a source, handle model uncertainty conservatively, and keep advisor sign-off, investment decisions, and output review with people. Because early-stage screening leans almost entirely on documents supplied by the prospect developer, information completeness, public-source cross-checking, and per-datum source referencing emerged as the non-negotiable reliability demands. Practitioners asked in addition for institutional memory, a logged record of the prospects offered and the reasons for rejection (FR-05, FR-18), and for a helicopter overview with drill-down for specialists (US-01). Elicited without presupposing a technology, the requirements describe an instrument that must operate natively on unstructured, largely textual material under explicit human control. Practice thereby corroborated, on independent grounds, the LLM direction that the knowledge gap of Chapter 2 had already indicated.

SQ4: How can the identified risk factors and requirements be realised in a working LLM-based prototype? By breaking the investment screening process into manageable, separately verifiable steps. The gain of that split is that each step can be specified, checked, and improved on its own, and that a failure can be located in the step that produced it rather than attributed to the system as a whole. The conceptual design therefore decomposes LLM-integrated prospect analysis into four blocks that distinguish three kinds of work: data retrieval (filling the checklist), knowledge updating (maintaining the market memory and the organisational assessment memory), and judgement (assessing the prospect). The three kinds have different correctness criteria and different input needs. Managing the input of each step separately is what bounds hallucination. A call that is asked only to retrieve, and that receives only the documents relevant to the items in front of it, has a narrower task to fail at and less unrelated material to reason from than a single call asked to read everything and conclude. Fabrication is most likely precisely where a model is asked to produce something the supplied material does not contain, so narrowing both the question and the input narrows the opportunity. The same separation keeps every output traceable to a process step that can be improved. The foundational retrieval block was implemented as the working prototype. Its three phases follow the information actually available as a prospect matures: a quick extraction from the teaser, a full extraction from the data room, and a comparison of selected findings against publicly available sources. Each phase ends in a human review gate where the expert confirms, corrects, or rejects findings, and rejection feedback is injected into the prompt of the re-run, so expert knowledge sharpens the extraction instead of being overwritten by it. The direction of that loop is what matters. A gate that only accepts or rejects makes the expert a checker of machine output, whereas a gate that returns the reason for a rejection to the next attempt makes the expert's knowledge part of the instrument itself, so review effort accumulates in the tool rather than being spent again at every run. The implementation confirmed two disciplines as necessary rather than optional. Precise specification of what the model must extract is a must, since each checklist item is compiled directly into the prompt; and requests must stay small and concise, with page budgets and adaptive batching keeping any single call within what can be answered reliably. The record itself enforces the requirements: 152 items, each with an evidence state (*found*, *probably found*, *not found*), a machine-checked source quote, and a cross-phase history in which human decisions are never silently overwritten. Once this structuring block is dependable, the two memories can be defined, and only then can a fully LLM-supported investment risk assessment be developed on top of them.

SQ5: How does the prototype perform under realistic screening conditions, and how do domain experts assess its fit with current practice? The expert evaluation returns a two-sided verdict: the prototype improves on current practice in structuring the screening record, while the evidence trail beneath that record does not yet meet the standard that due diligence requires of it. On the improvement side, nine real prospects were processed at negligible marginal API cost, producing standardised, source-attributed records where previously the same reading was repeated across specialists and nothing was systematically logged. The record and its screening opportunities (per-item provenance, cross-phase conflict detection, logged rejections) did not exist in current practice in any single environment. The independent public-source check corrected or extended a majority of the claims it verified, which supports the value of an independent leg against bottleneck B1, and consistency probing across near-identical sibling prospects classified the large majority of items identically, with the divergences traced mainly to under-specified items rather than to the model (Sections 8.2 and 8.3 give the figures). The practitioners endorsed both the design and its deliberate scope, and asked to start using the structuring layer.

On the evidence side, the in-depth expert walk-through delivered the evaluation's sharpest finding: the extraction was frequently right in substance, but it reached its answers through the seller's general overview documents rather than through the underlying evidence. In the audited run, 61% of quotations rested on documents composed by the selling party itself, where the auditing expert considers 5 to 10% adequate, and the cited documents covered a small fraction of the data room (Section 8.4). The record is therefore usable as a structured first pass, and not yet reliable as a due-diligence record. The distinction rests on what due diligence is for: its purpose is to test the seller's account against the underlying instruments, so a record whose evidence is largely the seller's own summary reproduces that account instead of verifying it. A value can be correct and still be unverified, and the record as it stands does not let the reader tell those two cases apart. The observed shortfall is consistent with the specification of the items and the composition of the documents the model receives, both properties of the design, although the evaluation did not isolate these from model behaviour. The prototype has a refinement backlog and is not yet operationally deployable, but it demonstrates that LLM-supported investment screening is achievable; the decisive adoption quantities (extraction accuracy and time saved against a known baseline) remain deliberately unmeasured by this formative evaluation and define the follow-up study.

Main research question

How can a Large Language Model-based screening instrument be designed to convert unstructured prospect documentation into a screening record that supports early-stage investment risk assessment for utility-scale BESS acquisition?

The answer this thesis gives is to split the work between structuring and judgement: give the LLM the structuring layer under explicit human gates, and govern it not by instantly trusting the model, but by designing what it is asked, what it is given, and how its every claim can be checked. With the right checks and balances, and with the screening process cut into manageable steps, LLMs can elevate the level of early-stage investment screening: more prospects can be screened with the same expert capacity, a currently unstructured process acquires structure, and data that was previously locked in documents and in employees' heads becomes accessible and comparable. This thesis demonstrates that concept under the tested configuration, and delivers what the full process requires around it: a catalogue of 52 risk factors that defines what an assessment must cover, 41 requirements that define how a supporting instrument must behave, a four-block process that defines where the LLM fits, and a working, formatively evaluated prototype of the foundational block that shows how it is built.

Control over output quality, moreover, is exercised largely through the design, and the reason is that each control removes a specific freedom from the model rather than asking it to behave well. The evidence states remove the option of answering without committing to how certain the answer is. The machine-checked citation removes the option of supporting a claim with a quote that does not appear in the cited document. Sticky ownership removes the option of overwriting a decision a person has already made. The spending caps remove the option of exceeding the budget at all, because they are enforced in the engine and not requested in the prompt. None of these mechanisms depends on the model's cooperation, and they therefore function as controls rather than as instructions.

The evaluation supports this reading from both sides. Where output fell short, the failures were traceable to the specificity of the instructions and to the composition of the documents placed before the model. And where the model was merely asked to prefer better sources, its behaviour changed little, while what it was given to read determined what it could cite at all. Because the evaluation varied neither model nor routing, it does not establish that design dominates model capability; what it does show is that the specification and the document set are the levers the designer can actually pull. Whether the same holds at the judgement layer is deliberately left untested here, because that layer was not built: it presupposes a dependable structuring layer beneath it, and its outputs lack the near-binary correctness that made it possible to attribute a shortfall to specification rather than to the model.

A demonstrated concept is, however, not a finished instrument. Production use requires the security hardening identified by a BESS developer's own review, a revisited confidentiality requirement (QL-06), and, given the sensitivity of acquisition material, serious consideration of running the model on infrastructure under the BESS developer's own control. The path beyond the prototype is equally concrete: complete the item specification, type the data-room documents by their author, build the two memories, and only then construct the assessment stage, with the known-baseline study (accuracy and time saved on completed prospects) as the decisive next measurement. The formative evidence supports this direction: practitioners endorsed the design and its scope, and asked to start using the structuring layer as the first step on that path.

Academic Contribution

Chapter 2 identified a threefold gap: no operationalised set of risk factors for early-stage BESS acquisition, no assessment approach calibrated to the conditions of the context, and no designed and evaluated LLM-based screening instrument for it. Confronting that gap with what has been delivered gives the thesis's academic contribution its shape.

The substantive gap is closed by the risk-factor catalogue. The infrastructure-risk literature offered category-level typologies (market, completion, and institutional risks in Miller and Lessard [68]; technical, economic, social, and institutional categories in Abba, Balta-Ozkan, and Hart [1]) but no factor-level, source-attributed inventory that a screening for this asset class could operationalise. The 52-factor catalogue extends those typologies to factor level for BESS acquisition, integrates the dispersed indicator-level material of sectoral and grey literature with practitioner knowledge, and preserves per-factor traceability to its sources. It is, to the author's knowledge, the first such catalogue for the utility-scale BESS context, and it is packaged for reuse as the judgement layer of future assessment instruments.

The methodological gap is closed in part and reframed in part. Section 2.4 showed that no single method family satisfies the four contextual conditions at once, and the reframing this thesis defends is that no single method has to: the conditions distribute over a layered process. The structuring layer demonstrated here meets three of them by construction. It operates on mixed, largely qualitative input; it needs no historical dataset, confirming in an unstudied setting the finding of Erfani and Khanjar [27] that competitive LLM extraction requires no task-specific labelled data; and it scales across a pipeline at negligible marginal cost. The fourth condition, representing interdependencies between risk factors, is deliberately assigned to the judgement layer, where the catalogue and its causal structure become the working material, and remains open. The evaluation also sharpens the hybrid division of labour reported by Martin, James, and Chadee [66], in which the model supplies breadth and the human supplies depth. It does so by specifying the concrete mechanisms of that division: evidence states that prefer an honest absence, per-item sources with machine-checked quotes, phase gates with sticky human ownership, and honesty rules governing what the model may claim. And it nuances the domain-adaptation thesis of the financial-LLM literature [58, 42]: in this design, domain knowledge enters through the instrument (the items and skills compiled into every prompt) rather than through the model's weights, and a general-purpose model sufficed for the extraction task.

The artefactual gap is closed with an instrument that is designed, evaluated, and documented: the complete task specification, prompt files, and payload anatomy are published in the appendices, which answers the standing critique that design-science artefacts are reported too opaquely to transfer [71]. The evaluation itself adds two methods to the toolbox for stochastic design artefacts: near-duplicate probing, which measures classification consistency without ground truth, and a repeated-request check,

which showed that temperature-zero decoding narrows but does not close run-to-run variance and thereby bounds what any single run can claim.

One contribution was not anticipated in the research design. The LLM engineering literature treats the provenance of a generated claim as a property of the model's output: models are made verifiable by requiring citations and checking them [34, 59], a framing this thesis initially adopted in its own provenance requirement (QL-01). The expert evaluation shows that framing to be insufficient when the document set is assembled by an interested party. A citation can name a real document and carry a verifiable quote and still fail its purpose, because the document is the seller's own summary of the evidence rather than the evidence: in the audited run, 61% of quotations rested on seller-composed material. Provenance quality in such settings is governed upstream of the model, by the composition of the documents placed before it; stated as a requirement, the difference is between naming a source and naming an admissible one. Beyond these findings, the thesis contributes prescriptive knowledge in the sense of Iivari [44]: eight design principles for integrating LLMs into confidential, high-stakes document work, distilled from a build that practitioners intend to adopt.

The research agenda that follows from these contributions is split by nature. The engineering continuation of the prototype, including the document-set ablation and the item-specification experiment that would sharpen the two central findings, is consolidated as a development roadmap in Appendix E; the broader academic directions this work opens (from decision-centred derivation of the screening information model to the assurance of AI on counterparty-confidential material) are developed in Section 9.8.

Societal Contribution

The societal case follows the chain established in Chapter 1. Electricity systems need storage faster than it is currently realised, and the pace of realisation depends in part on how quickly and how well viable projects are identified and funded. The theory of change this thesis works from is that an instrument raising the screening capacity of a BESS developer, and improving the evidential basis on which prospects are accepted or rejected, would raise that pace at the scale of each organisation that adopts it. Neither screening capacity nor decision error was measured here, so this is a stated expectation with testable assumptions rather than a demonstrated effect. The thesis also contributes a worked example of accountable LLM use on confidential, consequential material: human accountability is preserved by construction, every claim is traceable to its source, and the system records what it did not read, did not find, and did not verify. It is not yet a responsible production deployment, because the adversarial-input risks of Section 9.5 remain untested. As organisations decide how to admit LLMs into their core processes, and weigh public API access against models run on their own infrastructure, the governance reasoning documented here (Section 9.5) offers a reusable pattern that does not surrender verifiability. For the BESS developer and its peers, the recommendations are stated in order in Section 9.4: specify the checklist items before further model work, type documents by their author, and carry confirmed values into every later extraction step.

Link to the MSc Programme

The research is a Complex Systems Engineering and Management thesis in substance. The problem is socio-technical: the screening of BESS acquisitions sits at the intersection of a technical artefact class, a multi-actor institutional environment of prospect developers, grid operators, and permitting authorities, and an organisation's own decision processes. The intervention is a designed system in that environment: it required engineering (a working pipeline with measured behaviour), institutional analysis (bottlenecks, requirements, human accountability boundaries), and process design (the four-block decomposition and its gates) to be developed together. The thesis manages uncertainty the CoSEM way: by structuring information, recording its provenance, and placing expert judgement at defined review points rather than outside the system.

References

[1] Z. Y. I. Abba, N. Balta-Ozkan, and P. Hart. “A holistic risk management framework for renewable energy investments”. In: *Renewable and Sustainable Energy Reviews* 160 (May 2022), p. 112305. ISSN: 1364-0321. DOI: 10.1016/j.rser.2022.112305. URL: https://www.sciencedirect.com/science/article/pii/S1364032122002210 (visited on 03/17/2026).

[2] William Adams. “Conducting Semi-Structured Interviews”. In: Aug. 2015. ISBN: 978-1-118-89360-9. DOI: 10.1002/9781119171386.ch19.

[3] Agile Business Consortium. *What is MoSCoW Prioritization?* en-GB. May 2026. URL: https://www.agilebusiness.org/resource/what-is-moscow-prioritization/ (visited on 06/03/2026).

[4] Maxine Akhurst et al. “Storage Readiness Levels: communicating the maturity of site technical understanding, permitting and planning needed for storage operations using CO2”. In: *International Journal of Greenhouse Gas Control* 110 (Sept. 2021), p. 103402. ISSN: 1750-5836. DOI: 10.1016/j.ijggc.2021.103402. URL: https://www.sciencedirect.com/science/article/pii/S1750583621001547 (visited on 04/08/2026).

[5] Carol Alexander and Xi Chen. “Model risk in real option valuation”. en. In: *Annals of Operations Research* 299.1 (Apr. 2021), pp. 1025–1056. ISSN: 1572-9338. DOI: 10.1007/s10479-019-03273-4. URL: https://doi.org/10.1007/s10479-019-03273-4 (visited on 05/12/2026).

[6] Àlex Alonso-Travesset et al. “Economic and Regulatory Uncertainty in Renewable Energy System Design: A Review”. en. In: *Energies* 16.2 (Jan. 2023). ISSN: 1996-1073. DOI: 10.3390/en16020882. URL: https://www.mdpi.com/1996-1073/16/2/882 (visited on 04/23/2026).

[7] Anthropic. *Prompt caching*. 2026. URL: https://docs.claude.com/en/docs/build-with-claude/prompt-caching (visited on 07/07/2026).

[8] Neeraj Arora, Mark Lazaroff, and Thomas D’Ambrosio. *A 2025 Update on Utility-Scale Energy Storage Procurements*. en. Mar. 2025. URL: https://www.morganlewis.com/pubs/2025/03/a-2025-update-on-utility-scale-energy-storage-procurements (visited on 04/30/2026).

[9] Asian Development Bank. *Handbook on Battery Energy Storage System*. en. Tech. rep. Edition: 0. Manila, Philippines: Asian Development Bank, Dec. 2018. DOI: 10.22617/TCS189791-2. URL: https://www.adb.org/publications/battery-energy-storage-system-handbook (visited on 04/09/2026).

[10] Autoriteit Consument & Markt. *ACM has established measures against grid congestion and published basic principles for future tariff changes | ACM*. en. News. Last Modified: 2024-07-19T10:30:00+0200. July 2024. URL: https://www.acm.nl/en/publications/acm-has-established-measures-against-grid-congestion-and-published-basic-principles-future-tariff-changes (visited on 04/30/2026).

[11] Vivek Bhagwan, Sara Grobbelaar, and Wouter Bam. “A SYSTEMATIC REVIEW OF THE DUE DILIGENCE STAGE OF MERGERS AND ACQUISITIONS: TOWARDS a CONCEPTUAL FRAMEWORK”. In: *The South African Journal of Industrial Engineering* 29.3 (Nov. 2018). DOI: 10.7166/29-3-2061. URL: https://doi.org/10.7166/29-3-2061.

[12] Glenn A. Bowen. “Document Analysis as a Qualitative Research Method”. In: *Qualitative Research Journal* 9.2 (Aug. 2009), pp. 27–40. ISSN: 1443-9883. DOI: 10.3316/QRJ0902027. URL: https://doi.org/10.3316/QRJ0902027 (visited on 05/19/2026).

[13] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (Jan. 2006). _eprint: https://doi.org/10.1191/1478088706qp063oa, pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa. URL: https://doi.org/10.1191/1478088706qp063oa (visited on 05/19/2026).

[14] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (visited on 05/19/2026).

[15] John M Bryson. “What to do when Stakeholders matter”. In: *Public Management Review* 6.1 (Mar. 2004). _eprint: https://doi.org/10.1080/14719030410001675722, pp. 21–53. ISSN: 1471-9037. DOI: 10.1080/14719030410001675722. URL: https://doi.org/10.1080/14719030410001675722 (visited on 04/10/2026).

[16] Rubina Canesi et al. “Probabilistic risk assessment framework for cost overruns predictions in infrastructure projects using randomized simulations”. In: *Computer-Aided Civil and Infrastructure Engineering* 40.27 (Nov. 2025), pp. 4774–4796. ISSN: 1093-9687. DOI: 10.1111/mice.70100. URL: https://www.sciencedirect.com/science/article/pii/S1093968726015239 (visited on 03/17/2026).

[17] Chris Chapman and Stephen Ward. *Project risk management: processes, techniques and insights*. John Wiley & Sons, 1996. ISBN: 978-0-471-95804-8. URL: https://eprints.soton.ac.uk/35673/ (visited on 03/31/2026).

[18] Michele Chinosi and Alberto Trombetta. “BPMN: An introduction to the standard”. In: *Computer Standards & Interfaces* 34.1 (Jan. 2012), pp. 124–134. ISSN: 0920-5489. DOI: 10.1016/j.csi.2011.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0920548911000766 (visited on 05/13/2026).

[19] Wesley Cole, A Will Frazier, and Chad Augustine. *Cost Projections for Utility-Scale Battery Storage: 2021 Update*. en. Tech. rep. DE-AC36-08GO28308. Denver: National Renewable Energy Laboratory (USA), June 2021. URL: https://docs.nlr.gov/docs/fy21osti/79236.pdf.

[20] Nils Collath et al. “Aging aware operation of lithium-ion battery energy storage systems: A review”. In: *Journal of Energy Storage* 55 (Nov. 2022), p. 105634. ISSN: 2352-152X. DOI: 10.1016/j.est.2022.105634. URL: https://www.sciencedirect.com/science/article/pii/S2352152X2201622X (visited on 03/19/2026).

[21] John Creswell. *Research design: Qualitative, quantitative and mixed methods*. Sage Publications, 1994.

[22] Wassim Daher et al. “An analysis of mergers in the presence of uncertainty in renewable energy integration costs”. In: *Energy Economics* 150 (Oct. 2025), p. 108819. ISSN: 0140-9883. DOI: 10.1016/j.eneco.2025.108819. URL: https://www.sciencedirect.com/science/article/pii/S0140988325006462 (visited on 02/13/2026).

[23] Directorate-General for Energy of the EC. *Accelerate the rollout of renewable energy*. en. Text. Mar. 2023. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2061 (visited on 03/18/2026).

[24] Avinash K. Dixit and Robert S. Pindyck. *Investment Under Uncertainty*. en. Google-Books-ID: VahsELa_qC8C. Princeton University Press, 1994. ISBN: 978-0-691-03410-2.

[25] Florian Egli. “Renewable energy investment risk: An investigation of changes over time and the underlying drivers”. In: *Energy Policy* 140 (May 2020), p. 111428. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111428. URL: https://www.sciencedirect.com/science/article/pii/S0301421520301816 (visited on 03/25/2026).

[26] Dina A. Elalfy et al. “Comprehensive review of energy storage systems technologies, objectives, challenges, and future trends”. In: *Energy Strategy Reviews* 54 (July 2024), p. 101482. ISSN: 2211-467X. DOI: 10.1016/j.esr.2024.101482. URL: https://www.sciencedirect.com/science/article/pii/S2211467X24001895 (visited on 04/08/2026).

[27] Abdolmajid Erfani and Hussein Khanjar. “Large Language Models for Construction Risk Classification: A Comparative Study”. en. In: *Buildings* 15.18 (Sept. 2025). ISSN: 2075-5309. DOI: 10.3390/buildings15183379. URL: https://www.mdpi.com/2075-5309/15/18/3379 (visited on 05/19/2026).

[28] K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. Revised. Cambridge, MA: MIT Press, 1993.

[29] European Union Agency for the Cooperation of Energy Regulators. *Electricity infrastructure development to support a competitive and sustainable energy system*. Tech. rep. Dec. 2024. URL: https://www.acer.europa.eu/sites/default/files/documents/Publications/ACER_2024_Monitoring_Electricity_Infrastructure.pdf.

[30] Jana Fabianová et al. "A Comprehensive Methodology for Investment Project Assessment Based on Monte Carlo Simulation". en. In: *Applied Sciences* 13.10 (May 2023). ISSN: 2076-3417. DOI: 10.3390/app13106103. URL: https://www.mdpi.com/2076-3417/13/10/6103 (visited on 03/17/2026).

[31] Roberto Dias Filho et al. "Strategic Guidelines for Battery Energy Storage System Deployment: Regulatory Framework, Incentives, and Market Planning". en. In: *Energies* 16.21 (Oct. 2023). ISSN: 1996-1073. DOI: 10.3390/en16217272. URL: https://www.mdpi.com/1996-1073/16/21/7272 (visited on 04/08/2026).

[32] Bent Flyvbjerg. "Quality control and due diligence in project management: Getting decisions right by taking the outside view". In: *International Journal of Project Management* 31.5 (Nov. 2012), pp. 760–774. DOI: 10.1016/j.ijproman.2012.10.007. URL: https://doi.org/10.1016/j.ijproman.2012.10.007.

[33] Bent Flyvbjerg and Dan Gardner. *How Big Things Get Done: The Surprising Factors That Determine the Fate of Every Project, from Home Renovations to Space Exploration and Everything In Between*. en. Google-Books-ID: uuRzEAAAQBAJ. Crown, Feb. 2023. ISBN: 978-0-593-23952-0.

[34] Tianyu Gao et al. "Enabling Large Language Models to Generate Text with Citations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6465–6488. DOI: 10.18653/v1/2023.emnlp-main.398. URL: https://aclanthology.org/2023.emnlp-main.398/.

[35] Nadine Gatzert and Nikolai Vogl. "Evaluating investments in renewable energy under policy risks". In: *Energy Policy* 95 (Aug. 2016), pp. 238–252. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2016.04.027. URL: https://www.sciencedirect.com/science/article/pii/S0301421516301951 (visited on 03/25/2026).

[36] Zeenat Hameed and Chresten Træholt. *EPC Framework for BESS Projects*. arXiv:2506.13281 [eess] version: 1. June 2025. DOI: 10.48550/arXiv.2506.13281. URL: http://arxiv.org/abs/2506.13281 (visited on 04/08/2026).

[37] Zeenat Hameed et al. "A business-oriented approach for battery energy storage placement in power systems". In: *Applied Energy* 298 (Sept. 2021), p. 117186. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.117186. URL: https://www.sciencedirect.com/science/article/pii/S0306261921006140 (visited on 02/17/2026).

[38] Ali Hashemizadeh et al. "Renewable energy investment risk assessment in belt and road initiative countries under uncertainty conditions". In: *Energy* 214 (Jan. 2021), p. 118923. ISSN: 0360-5442. DOI: 10.1016/j.energy.2020.118923. URL: https://www.sciencedirect.com/science/article/pii/S0360544220320302 (visited on 04/08/2026).

[39] Alan R. Hevner. "A Three Cycle View of Design Science Research". In: *Scandinavian Journal of Information Systems* 19.2 (2007), pp. 87–92. URL: https://aisel.aisnet.org/sjis/vol19/iss2/4.

[40] Alan R. Hevner et al. "Design Science in Information Systems Research1". en. In: *Management Information Systems Quarterly* 28.1 (Mar. 2004), pp. 75–106. ISSN: 0276-7783. DOI: 10.2307/25148625. URL: https://dx.doi.org/10.2307/25148625 (visited on 02/11/2026).

[41] Yu Hu, Miguel Armada, and María Jesús Sánchez. "Potential utilization of battery energy storage systems (BESS) in the major European electricity markets". In: *Applied Energy* 322 (June 2022), p. 119512. DOI: 10.1016/j.apenergy.2022.119512. URL: https://doi.org/10.1016/j.apenergy.2022.119512.

[42] Allen H. Huang, Hui Wang, and Yi Yang. “FinBERT: A Large Language Model for Extracting Information from Financial Text”. en. In: *Contemporary Accounting Research* 40.2 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3846.12832, pp. 806–841. ISSN: 1911-3846. DOI: 10.1111/1911-3846.12832. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12832 (visited on 05/19/2026).

[43] Georgina Hutton and Iona Stewart. *Battery energy storage systems Research Briefing*. en-GB. Tech. rep. 7621. House of Commons Library, June 2025. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-7621/ (visited on 04/30/2026).

[44] Juhani Iivari. “Distinguishing and contrasting two strategies for design science research”. In: *European Journal of Information Systems* 24.1 (Jan. 2015). _eprint: https://doi.org/10.1057/ejis.2013.35, pp. 107–115. ISSN: 0960-085X. DOI: 10.1057/ejis.2013.35. URL: https://doi.org/10.1057/ejis.2013.35 (visited on 03/18/2026).

[45] Intergovernmental Panel on Climate Change, WMO, and UNEP. *Climate Change 2022: mitigation of climate change*. eng. Num Pages: 2042. Geneva, 2022. URL: https://digitallibrary.un.org/record/4019730 (visited on 04/23/2026).

[46] International Energy Agency. *Battery storage is scaling up and taking on a larger system role*. 2026. URL: https://www.iea.org/commentaries/battery-storage-is-scaling-up-and-taking-on-a-larger-system-role (visited on 07/06/2026).

[47] International Energy Agency. *Energy Policy Review The Netherlands 2024*. en-GB. Jan. 2025. URL: https://www.iea.org/reports/the-netherlands-2024 (visited on 04/30/2026).

[48] International Organization for Standardization and International Electrotechnical Commission. “ISO/IEC/IEEE International Standard - Systems and software engineering – Life cycle processes – Requirements engineering”. In: *ISO/IEC/IEEE 29148:2018(E)* 2 (Nov. 2018), pp. 1–104. DOI: 10.1109/IEEESTD.2018.8559686. URL: https://ieeexplore.ieee.org/document/8559686 (visited on 06/04/2026).

[49] International Renewable Energy Agency. *World Energy Transitions Outlook 2023: 1.5°C Pathway*. en. Tech. rep. Abu Dhabi., June 2023. URL: https://www.irena.org/Publications/2023/Jun/World-Energy-Transitions-Outlook-2023 (visited on 04/30/2026).

[50] Anastasia Ioannou, Andrew Angus, and Feargal Brennan. “Risk-based methods for sustainable energy system planning: A review”. In: *Renewable and Sustainable Energy Reviews* 74 (July 2017), pp. 602–615. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.02.082. URL: https://www.sciencedirect.com/science/article/pii/S136403211730312X (visited on 03/25/2026).

[51] Muhammad Saiful Islam et al. “A knowledge-based expert system to assess power plant project cost overrun risks”. In: *Expert Systems with Applications* 136 (Dec. 2019), pp. 12–32. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.06.030. URL: https://www.sciencedirect.com/science/article/pii/S0957417419304300 (visited on 03/17/2026).

[52] Xiao-Hua Jin and Guomin Zhang. “Modelling optimal risk allocation in PPP projects using artificial neural networks”. In: *International Journal of Project Management* 29.5 (July 2011), pp. 591–603. ISSN: 0263-7863. DOI: 10.1016/j.ijproman.2010.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0263786310001158 (visited on 03/18/2026).

[53] Paul Johannesson and Erik Perjons. *An Introduction to Design Science*. en. Cham: Springer International Publishing, 2014. ISBN: 978-3-319-10631-1. DOI: 10.1007/978-3-319-10632-8. URL: https://link.springer.com/10.1007/978-3-319-10632-8 (visited on 02/11/2026).

[54] Amirhossein Karamoozian et al. “Risk assessment of renewable energy projects using uncertain information”. en. In: *International Journal of Energy Research* 46.13 (2022). _eprint: https://onlinelibrary.wiley.com pp. 18079–18099. ISSN: 1099-114X. DOI: 10.1002/er.8428. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/er.8428 (visited on 03/24/2026).

[55] Salim Ullah Khan et al. “A Data-Driven Bayesian Belief Network Influence Diagram Approach for Socio-Environmental Risk Assessment and Mitigation in Major Ecosystem- and Landscape-Modifier Projects”. en. In: *Sustainability* 17.8 (Apr. 2025). ISSN: 2071-1050. DOI: 10.3390/su17083537. URL: https://www.mdpi.com/2071-1050/17/8/3537 (visited on 03/17/2026).

[56] Marvin Killer, Mana Farrokhseresht, and Nikolaos G. Paterakis. “Implementation of large-scale Li-ion battery energy storage systems within the EMEA region”. In: *Applied Energy* 260 (Feb. 2020), p. 114166. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.114166. URL: https://www.sciencedirect.com/science/article/pii/S0306261919318537 (visited on 04/08/2026).

[57] Chun Sing Lai and Giorgio Locatelli. “Economic and financial appraisal of novel large-scale energy storage technologies”. In: *Energy* 214 (Oct. 2020), p. 118954. DOI: 10.1016/j.energy.2020.118954. URL: https://doi.org/10.1016/j.energy.2020.118954.

[58] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. “Large Language Models in Finance (FinLLMs)”. en. In: *Neural Computing and Applications* 37.30 (Oct. 2025), pp. 24853–24867. ISSN: 1433-3058. DOI: 10.1007/s00521-024-10495-6. URL: https://doi.org/10.1007/s00521-024-10495-6 (visited on 05/19/2026).

[59] Nelson F. Liu, Tianyi Zhang, and Percy Liang. “Evaluating Verifiability in Generative Search Engines”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, Dec. 2023. DOI: 10.18653/v1/2023.findings-emnlp.467. URL: https://aclanthology.org/2023.findings-emnlp.467/.

[60] Nelson F. Liu et al. “Lost in the Middle: How Language Models Use Long Contexts”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173. DOI: 10.1162/tacl_a_00638. URL: https://aclanthology.org/2024.tacl-1.9/.

[61] Wei Liu et al. “Consensus-based multidimensional due diligence of fintech-enhanced green energy investment projects”. In: *Financial Innovation* 7.1 (Dec. 2021). DOI: 10.1186/s40854-021-00289-3. URL: https://doi.org/10.1186/s40854-021-00289-3.

[62] Yi Liu et al. “Prospect Theory-Based q-Rung Orthopair Fuzzy TODIM Method for Risk Assessment of Renewable Energy Projects”. en. In: *International Journal of Fuzzy Systems* 26.3 (Apr. 2024), pp. 1046–1068. ISSN: 2199-3211. DOI: 10.1007/s40815-023-01652-5. URL: https://doi.org/10.1007/s40815-023-01652-5 (visited on 03/17/2026).

[63] Giorgio Locatelli, Diletta Colette Invernizzi, and Mauro Mancini. “Investment and risk appraisal in energy storage systems: A real options approach”. In: *Energy* 104 (June 2016), pp. 114–131. ISSN: 0360-5442. DOI: 10.1016/j.energy.2016.03.098. URL: https://www.sciencedirect.com/science/article/pii/S0360544216303425 (visited on 04/08/2026).

[64] Xing Luo et al. “Overview of current development in electrical energy storage technologies and the application potential in power system operation”. In: *Applied Energy* 137 (Jan. 2015), pp. 511–536. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2014.09.081. URL: https://www.sciencedirect.com/science/article/pii/S0306261914010290 (visited on 03/12/2026).

[65] Subir Majumder et al. “Exploring the capabilities and limitations of large language models in the electric energy sector”. English. In: *Joule* 8.6 (June 2024), pp. 1544–1549. ISSN: 2542-4785, 2542-4351. DOI: 10.1016/j.joule.2024.05.009. URL: https://www.cell.com/joule/abstract/S2542-4351(24)00237-X (visited on 05/19/2026).

[66] Hector Martin, Jennifer James, and Aaron Chadee. “Exploring Large Language Model AI tools in Construction Project Risk Assessment: Chat GPT Limitations in Risk Identification, Mitigation Strategies, and User Experience”. en. In: *Journal of Construction Engineering and Management* 151.9 (Sept. 2025), p. 04025119. DOI: 10.1061/JCEMD4.COENG-16658. URL: https://ascelibrary.org/doi/10.1061/JCEMD4.COENG-16658 (visited on 05/19/2026).

[67] Charilaos Mertzanis. “Artificial intelligence and investment management: Structure, strategy, and governance”. In: *International Review of Financial Analysis* 107 (Nov. 2025), p. 104599. ISSN: 1057-5219. DOI: 10.1016/j.irfa.2025.104599. URL: https://www.sciencedirect.com/science/article/pii/S1057521925006866 (visited on 02/13/2026).

[68] Roger Miller and Donald Lessard. “Understanding and managing risks in large engineering projects”. In: *International Journal of Project Management* 19.8 (Nov. 2001), pp. 437–443. ISSN: 0263-7863. DOI: 10.1016/S0263-7863(01)00045-X. URL: https://www.sciencedirect.com/science/article/pii/S026378630100045X (visited on 04/23/2026).

[69] Zachary Munn et al. “Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach”. In: *BMC Medical Research Methodology* 18.1 (2018), p. 143. ISSN: 1471-2288. DOI: 10.1186/s12874-018-0611-x.

[70] Benjamin Murgas, Alvin Henao, and Luceny Guzman. “Evaluation of Investments in Wind Energy Projects, under Uncertainty. State of the Art Review”. en. In: *Applied Sciences* 11.21 (Oct. 2021). ISSN: 2076-3417. DOI: 10.3390/app112110213. URL: https://www.mdpi.com/2076-3417/11/21/10213 (visited on 03/17/2026).

[71] Tadhg Nagle et al. *The Research Method we Need or Deserve? A Literature Review of the Design Science Research Landscape*. June 2020. DOI: 10.31219/osf.io/8t7ra. URL: https://osf.io/8t7ra_v1/ (visited on 03/18/2026).

[72] Maria Elena Nenni et al. “How artificial intelligence will transform project management in the age of digitization: a systematic literature review”. en. In: *Management Review Quarterly* 75.2 (June 2025), pp. 1669–1716. ISSN: 2198-1639. DOI: 10.1007/s11301-024-00418-z. URL: https://doi.org/10.1007/s11301-024-00418-z (visited on 02/18/2026).

[73] NFPA. *Standard for the Installation of Stationary Energy Storage Systems*. Oct. 2025. URL: https://docinfofiles.nfpa.org/files/AboutTheCodes/855/Errata_855_26_1.pdf.

[74] Nordic Energy Research and Nordic Council of Ministers. *The Project Development Process*. en. Tech. rep. 4. Nov. 2016, p. 14. URL: https://www.nordicenergy.org/app/uploads/2017/05/Guide-report-4-The-project-development-process.pdf.

[75] Olufemi Ogunniran et al. “Risk-Based Optimization of Renewable Energy Investment Portfolios: A Multi-Stage Stochastic Approach to Address Uncertainty”. en. In: *Applied Sciences* 15.5 (Feb. 2025). ISSN: 2076-3417. DOI: 10.3390/app15052346. URL: https://www.mdpi.com/2076-3417/15/5/2346 (visited on 03/17/2026).

[76] OWASP Foundation. *LLM Prompt Injection Prevention Cheat Sheet*. 2025. URL: https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html (visited on 07/07/2026).

[77] Meng Qi et al. “Assessing and mitigating potential hazards of emerging grid-scale electrical energy storage systems”. In: *Process Safety and Environmental Protection* 149 (May 2021), pp. 994–1016. ISSN: 0957-5820. DOI: 10.1016/j.psep.2021.03.042. URL: https://www.sciencedirect.com/science/article/pii/S0957582021001634 (visited on 04/08/2026).

[78] Neil Rasmussen and Suzanne Niles. “Data Center Projects: Standardized Process”. en. In: WP140 Rev 2007-0 (2007). URL: http://media.comtec.com/pdf/apc/StandardisedProcess.pdf.

[79] Jan Recker. “Opportunities and constraints: the current struggle with BPMN”. In: *Business Process Management Journal* 16.1 (Feb. 2010), pp. 181–201. ISSN: 1463-7154. DOI: 10.1108/14637151011018001. URL: https://doi.org/10.1108/14637151011018001 (visited on 05/13/2026).

[80] Matthew Renze. “The Effect of Sampling Temperature on Problem Solving in Large Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7346–7356. DOI: 10.18653/v1/2024.findings-emnlp.432. URL: https://aclanthology.org/2024.findings-emnlp.432/ (visited on 07/12/2026).

[81] Paulo Rotella Junior et al. “Economic Analysis of the investments in battery energy Storage Systems: Review and Current Perspectives”. In: *Energies* 14.9 (Apr. 2021), p. 2503. DOI: 10.3390/en14092503. URL: https://doi.org/10.3390/en14092503.

[82] Abdulazeez Y. H. Saif-Alyousfi and Turki Rashed Alshammari. “Artificial intelligence adoption and private investment in renewable energy: Evidence from firm-level data across developed and emerging economies”. In: *Energy Strategy Reviews* 64 (Mar. 2026), p. 102054. ISSN: 2211-467X. DOI: 10.1016/j.esr.2026.102054. URL: https://www.sciencedirect.com/science/article/pii/S2211467X26000209 (visited on 02/13/2026).

[83] Benjamin Saunders et al. “Saturation in qualitative research: exploring its conceptualization and operationalization”. en. In: *Quality & Quantity* 52.4 (July 2018), pp. 1893–1907. ISSN: 1573-7845. DOI: 10.1007/s11135-017-0574-8. URL: https://doi.org/10.1007/s11135-017-0574-8 (visited on 05/19/2026).

[84] Andrea Senova, Alica Tobisova, and Robert Rozenberg. “New Approaches to Project Risk Assessment Utilizing the Monte Carlo Method”. en. In: *Sustainability* 15.2 (Jan. 2023). ISSN: 2071-1050. DOI: 10.3390/su15021006. URL: https://www.mdpi.com/2071-1050/15/2/1006 (visited on 03/17/2026).

[85] Anuja Shaktawat and Shelly Vadhera. “Risk management of hydropower projects for sustainable development: a review”. en. In: *Environment, Development and Sustainability* 23.1 (Jan. 2021), pp. 45–76. ISSN: 1573-2975. DOI: 10.1007/s10668-020-00607-2. URL: https://doi.org/10.1007/s10668-020-00607-2 (visited on 03/24/2026).

[86] Ian Varela Soares, Romain Mauger, and Thauan Santos. “Considerations for benefit stacking policies in the EU electricity storage market”. In: *Energy Policy* 172 (Nov. 2022), p. 113333. DOI: 10.1016/j.enpol.2022.113333. URL: https://doi.org/10.1016/j.enpol.2022.113333.

[87] SolarPower Europe. *European Battery Market Outlook 2026–2030*. 2026. URL: https://www.solarpowereurope.org/insights/outlooks/european-battery-market-outlook-2026-2030-1 (visited on 07/06/2026).

[88] Robert Springer. *Framework for Project Development in the Renewable Energy Sector*. en. Tech. rep. NREL/TP–7A40-57963, 1064514. Feb. 2013, NREL/TP–7A40–57963, 1064514. DOI: 10.2172/1064514. URL: https://www.osti.gov/servlets/purl/1064514/ (visited on 04/09/2026).

[89] Marco Stecca et al. “A Comprehensive Review of the Integration of Battery Energy Storage Systems Into Distribution Networks”. In: *IEEE Open Journal of the Industrial Electronics Society* 1 (2020), pp. 46–65. ISSN: 2644-1284. DOI: 10.1109/OJIES.2020.2981832. URL: https://ieeexplore.ieee.org/document/9040552 (visited on 03/19/2026).

[90] Bo Sun et al. “An optimal sequential investment decision model for generation-side energy storage projects in China considering policy uncertainty”. In: *Journal of Energy Storage* 83 (Apr. 2024), p. 110748. ISSN: 2352-152X. DOI: 10.1016/j.est.2024.110748. URL: https://www.sciencedirect.com/science/article/pii/S2352152X24003323 (visited on 03/31/2026).

[91] Sustainable Energy Authority of Ireland. *Battery Storage Manual of Consenting Procedures*. Tech. rep. IE000527. Apr. 2025. URL: https://singlepointofcontact.seai.ie/sites/spoc/files/2025-11/Battery-Storage-Manual-SEAI.pdf.

[92] Elham Tabassi. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. en. Tech. rep. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology (U.S.), Jan. 2023, NIST AI 100–1. DOI: 10.6028/NIST.AI.100-1. URL: http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (visited on 06/03/2026).

[93] Osman Taylan et al. “Construction projects selection and risk assessment by fuzzy AHP and fuzzy TOPSIS methodologies”. en. In: *Applied Soft Computing* 17 (Apr. 2014), pp. 105–116. ISSN: 15684946. DOI: 10.1016/j.asoc.2014.01.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1568494614000040 (visited on 03/24/2026).

[94] The World Bank and PPIAF. *Guidelines to implement battery energy storage systems under public-private partnership structures*. Tech. rep. Jan. 2023. URL: https://documents1.worldbank.org/curated/en/099536501202316060/pdf/IDU0edcfc32c0825f040f509c0b0bbf49294e569.pdf.

[95] TNO. *Queue on the power grid could slow down the mobility transition*. en. Nov. 2025. URL: https://www.tno.nl/en/newsroom/2025/11/queue-power-grid-mobility-transition/ (visited on 01/18/2026).

[96] Kingsley Ukoba et al. “Optimizing renewable energy systems through artificial intelligence: Review and future prospects”. EN. In: *Energy & Environment* 35.7 (Nov. 2024), pp. 3833–3879. ISSN: 0958-305X. DOI: 10.1177/0958305X241256293. URL: https://doi.org/10.1177/0958305X241256293 (visited on 03/18/2026).

[97] UL. *9450A Standard for safety: Test Method for Battery Energy Storage Systems*. Mar. 2025. URL: https://ulnormdoc.com/wp-content/preview/UL%209540A%202025.pdf.

[98] John Venable, Jan Pries-Heje, and Richard Baskerville. “A Comprehensive Framework for Evaluation in Design Science Research”. en. In: *Design Science Research in Information Systems. Advances in Theory and Practice*. Ed. by Ken Peffers, Marcus Rothenberger, and Bill Kuechler. Berlin, Heidelberg: Springer, 2012, pp. 423–438. ISBN: 978-3-642-29863-9. DOI: 10.1007/978-3-642-29863-9_31.

[99] Paul Verhagen, Jing Hu, and Robert Harmsen. “Evaluating non-firm grid connections for battery energy storage systems: A co-optimization case study of the Netherlands”. In: *Energy Policy* 208 (Oct. 2025), p. 114903. DOI: 10.1016/j.enpol.2025.114903. URL: https://doi.org/10.1016/j.enpol.2025.114903.

[100] Geoff Walsham. “Doing Interpretive Research”. In: *European Journal of Information Systems* 15.3 (2006), pp. 320–330. DOI: 10.1057/palgrave.ejis.3000589.

[101] Qiang Wang, Yuanfan Li, and Rongrong Li. “Integrating artificial intelligence in energy transition: A comprehensive review”. In: *Energy Strategy Reviews* 57 (Jan. 2025), p. 101600. ISSN: 2211-467X. DOI: 10.1016/j.esr.2024.101600. URL: https://www.sciencedirect.com/science/article/pii/S2211467X24003092 (visited on 03/18/2026).

[102] Daniel Wangerin. “M&A Due Diligence, Post□Acquisition Performance, and Financial Reporting for Business Combinations”. In: *Contemporary Accounting Research* 36.4 (May 2019), pp. 2344–2378. DOI: 10.1111/1911-3846.12520. URL: https://doi.org/10.1111/1911-3846.12520.

[103] Brian Warshay. *Energy Storage Capacity Warranties: Beyond the Fine Print*. en. Apr. 2021. URL: https://www.dnv.com/article/energy-storage-capacity-warranties-beyond-the-fine-print-200339/ (visited on 04/30/2026).

[104] Philipp Winder, Christian Hildebrand, and Jochen Hartmann. “Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors”. en. In: *PLOS ONE* 20.6 (June 2025), e0325459. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0325459. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0325459 (visited on 05/19/2026).

[105] Sumanth Yamujala et al. “Multi-service based economic valuation of grid-connected battery energy storage systems”. In: *Journal of Energy Storage* 52 (Apr. 2022), p. 104657. DOI: 10.1016/j.est.2022.104657. URL: https://doi.org/10.1016/j.est.2022.104657.

[106] Qinghua Yu. “Constructing a Design Research Model—Exploration of Research Paradigms Based on Research Tree Diagram”. en. In: *Methodology Insight* 1.1 (Apr. 2025), pp. 35–49. ISSN: 3107-2135. DOI: 10.71290/mi01010003. URL: https://www.scholepress.hk/journal/mi/1/1/7 (visited on 03/18/2026).

[107] Tianlei Zang et al. “Integrated Planning and Operation Dispatching of Source–Grid–Load–Storage in a New Power System: A Coupled Socio–Cyber–Physical Perspective”. en. In: *Energies* 17.12 (June 2024). ISSN: 1996-1073. DOI: 10.3390/en17123013. URL: https://www.mdpi.com/1996-1073/17/12/3013 (visited on 04/08/2026).

[108] Yuyang Zeng et al. “Long-Duration Energy Storage: A Critical Enabler for Renewable Integration and Decarbonization”. en. In: *Energies* 18.3 (Jan. 2025). ISSN: 1996-1073. DOI: 10.3390/en18030466. URL: https://www.mdpi.com/1996-1073/18/3/466 (visited on 03/18/2026).

[109] Haoran Zhao, Sen Guo, and Huiru Zhao. “Comprehensive assessment for battery energy storage systems based on fuzzy-MCDM considering risk preferences”. In: *Energy* 168 (Feb. 2019), pp. 450–461. ISSN: 0360-5442. DOI: 10.1016/j.energy.2018.11.129. URL: https://www.sciencedirect.com/science/article/pii/S0360544218323405 (visited on 04/08/2026).

[110] Xiaowen Zhao et al. “Fuzzy Analytic Hierarchy Process-Based Investment Risk Evaluation for Infrastructure Construction Projects”. en. In: *Buildings* 15.5 (Feb. 2025). ISSN: 2075-5309. DOI: 10.3390/buildings15050756. URL: https://www.mdpi.com/2075-5309/15/5/756 (visited on 03/17/2026).

A

AI Statement

Artificial intelligence tools, LLMs in particular, were used throughout this research. Rather than asserting a binary between writing and not writing, this appendix discloses per activity what was supplied to the tools, what they produced, which of their output was retained, how it was reviewed, and which tasks were kept outside their scope.

Text Refinement

Grammarly was used to assist with language refinement (grammar, spelling, and readability). ChatGPT and Claude were used to obtain structural feedback and to clarify the phrasing of arguments during drafting and revision. In the final revision stage, Claude was additionally used in a structured feedback-integration workflow to help incorporate the supervisors' and reviewers' comments into the manuscript and to check terminology and numerical consistency across chapters, with every proposed change reviewed and accepted, adapted, or rejected by the author. These tools were used to improve the expression of ideas that were already the author's own; they were not used to generate the substantive content of the thesis.

Literature Search

Consensus.app was used to support the initial exploration and broad screening of academic literature across several research fields, which made it more efficient to identify potentially relevant studies and recurring thematic patterns. Following this AI-supported screening phase, all further literature selection, critical appraisal, synthesis, and interpretation were carried out independently by the author. The author manually reviewed and assessed each publication to ensure its relevance, methodological quality, and academic reliability.

Prototype Development

The prototype described in Chapter 7 would not have reached its current level without the help of artificial intelligence. The author had a basic working knowledge of programming and of Python, but could not have developed a prototype of this complexity unaided. Anthropic's models were used since GIGA's company policy approved processing data through their LLMs. Claude Code acted as a coding assistant that worked alongside the author on the evolving codebase. Several models were used for different purposes: Haiku 4.5 for debugging and small fixes, Sonnet 4.6 for reworking minor process arrangements, and Opus 4.8 for integrating larger components such as the data-room triage-and-routing mechanism.

The prototype was developed iteratively and under the author's direction. The author specified the process to be implemented and designed the intended front end and back end of the user interface, and then built the prototype outward from that design, starting from the application environment provided by Vondr.ai (with which GIGA Storage was running an AI pilot at the time). Development proceeded step by step: the author examined the front end from the user's perspective, evaluated how it behaved, and reported bugs and desired changes to Claude Code. Beyond the interface, the author also examined

the back end and drove improvements there. One example illustrates why this back-end scrutiny mattered. When asked to read the prospect PDFs and send them to the API, Claude Code did not send the documents themselves; it silently applied a local text-extraction step and sent only the extracted text. Discovering and correcting this kind of behaviour was possible only because the author inspected how the system actually worked rather than trusting that the instruction had been followed as intended.

Boundary of Independent Authorship

All substantive research decisions were made by the author: the formulation of the research question, the theoretical positioning, the research design, the selection of methods, the interpretation of the literature, the design of the process, and the direction of the prototype's development.

On the empirical material the boundary runs as follows. AI tools did not create interview or evaluation data and did not make final interpretations. They produced the automatic transcriptions of the recordings and assisted the preliminary organisation of the author's own session notes into draft structures. Every such draft was checked against the recordings and the source documents and independently revised by the author, and the codes, themes, bottlenecks, risk factors, and requirements were determined by the author.

On the written text, the retained output is limited to phrasing and structural suggestions that the author reviewed sentence by sentence and accepted, adapted, or rejected. On the prototype, generated code was retained where the author had specified the behaviour, inspected the result, and tested it against real prospect material, as described above.

The author takes full responsibility for the content, accuracy, and originality of this MSc thesis.

B

Interview Materials

B.1. Overview of the Conducted Interviews

To support Sub-Question 1, 2 and 3, nine semi-structured expert interviews were conducted between 15 April and 1 May 2026. The interview sample was assembled to cover the full spectrum of expertise involved in the early-stage investment risk assessment of utility-scale battery energy storage system (BESS) projects, ranging from business development to financial close and from internal staff to external advisors.

The interviews followed a semi-structured format, in which the interview protocol described in Section B.2 was used as a common analytical backbone, while the questioning was adapted to the specific expertise of each interviewee. This methodological choice follows the recommendations of [2]: the protocol ensured comparability across cases on the core themes, while the flexibility of the semi-structured format allowed each interviewee to elaborate on the topics most relevant to their professional perspective. As a consequence, certain questions were asked in detail to some interviewees and only briefly to others, depending on whether they were professionally positioned to comment substantively.

Each interview lasted approximately forty to fifty-five minutes (with the exception of one interview at thirty minutes), was conducted via Microsoft Teams or in person, and was recorded with informed consent. Transcripts were generated from the recordings and subsequently structured into thematic analyses. A member-checking step was performed for each interview, in which the interviewee received a summary of the conversation and was invited to flag factual inaccuracies. The summaries that follow in Section B.3 represent the member-checked versions.

Table B.1 provides an overview of the interview sample, including each interviewee's anonymised code, role, and interview date.

Table B.1: Overview of the conducted interviews.

Code	Role	Date
P1	Development Expert	15 April 2026
P2	Senior Attorney	15 April 2026
P3	Technical Consultant	20 April 2026
P4	Technical Consultant	20 April 2026
P5	Development Expert	22 April 2026
P6	Investment Expert	24 April 2026
P7	Engineering Consultant	29 April 2026
P8	Finance Expert	29 April 2026
P9	Legal Consultant	1 May 2026

The sample includes seven internal interviewees of the BESS developer, together covering business development, legal, technical, engineering, investment, and project finance expertise, and two external

advisors representing technical due diligence and legal due diligence respectively. This composition was selected to ensure that the empirical material captures both the internal decision-making perspective of the BESS developer and the comparative cross-client perspective of external advisors who work with multiple BESS developers.

B.2. Interview Protocol

The interview protocol below served as the analytical backbone for all nine interviews. In line with the semi-structured format, the interviewer did not read the questions literally, but used them as anchors and adapted the order and depth of questioning to each interviewee's expertise.

Current practice and bottlenecks

Q0. Could you briefly describe your role and how it relates to the screening and acquisition of new BESS projects?

Q1. How is risk assessment of new early-stage BESS project opportunities currently conducted? Could you walk me through what happens step by step when a new project arrives?

Q2. What are the main bottlenecks or pain points you experience in the current screening process?

Information gaps and opportunities

Q3. What information that is already available during the screening phase is currently underused — for example because it is difficult to interpret, scattered across documents, or hard to compare across projects?

Q4. Looking at the data and documents available during early screening — which critical insights could in your view be more systematically extracted or structured, for instance with the help of data analysis tools or artificial intelligence?

Risk assessment per dimension

Q5. When a project has an ongoing grid connection application with the TSO or DSO, how do you assess the likelihood that the connection will be granted? Which specific factors or indicators do you look at?

Q6. When a project has an ongoing or planned permitting procedure, how do you assess the risk that the permit will be delayed or refused? Which signals do you rely on?

Q7. Which indicators tell you that the land situation of a project is safe or problematic? What do you look at in an early stage?

Q8. Beyond Power, Plot, Permit, and Profit — are there other risk dimensions that you consider critical but that are not always systematically captured in the current screening process?

Requirements for the decision support tool

Q9. If you could design a tool yourself to support early-stage project screening, what should it show or do for you?

Q10. Which key performance indicators or risk metrics should such a tool absolutely include in order to be credible in your eyes?

Q11. How important is it to you that the tool makes its reasoning transparent — that you can see why it assigns a particular risk level?

Q12. What would convince you to actually use such a tool in your daily work? And what would make you stop using it?

Closing

Q13. Is there anything about the screening process, the risk assessment, or what such a tool should do that we have not discussed but that you consider important?

Q14. Is there anyone else, internally or externally, whom you would recommend speaking to about this topic?

B.3. Conducted Interview Summaries

The summaries in this section reflect the member-checked content of each interview. Each subsection begins with the interviewee's role and the date of the interview. The structure within each summary follows the thematic logic that emerged from the conversation, organised under sub-subsection headings.

Interview P1: development expert

Date: 15 April 2026

Current screening process

New project opportunities arrive at the BESS developer through an information memorandum from the selling party. Internally, the location, public sources (TSO network maps, the cadastre, zoning plans) and strategic fit are assessed first. If the project is of interest, a non-binding offer is submitted, after which a data room is opened. Due diligence is conducted both internally (business development, technical, legal) and externally (legal and technical advisors) and results in a report structured per 4P dimension, with risk flags, severity, and mitigation measures. This report goes to the Investment Committee for the final investment decision. The full DD process costs approximately €50,000–€100,000 per project and takes around four weeks of external advisor work.

Main bottlenecks

The high cost and lead time of the DD process are the primary pain points. Many manual checks are performed (land ownership, connection agreement, contracts). One missing element can stop an entire project — the example of a cable route agreement with a single landowner illustrates this. The earlier risks are identified, the lower the costs that must be incurred. Projects are informally benchmarked against earlier experience, but this is not systematically recorded; knowledge about rejected projects resides in employees' heads.

Risk assessment per dimension

For Power, the BESS developer does not enter a project without at least an offer or realisation agreement from the TSO. Network maps and TDTR contract status are key indicators. Typical connection costs for 100–300 MW range between €2–5 million; a 6 km cable can already cost €20 million. For Permit, the principebesluit (NL) or Vorbescheid (DE) is the most important early indicator; permitting risk is considered the largest uncertainty with the least control. Objections from residents are often not factually substantiated but can lead to lengthy and costly procedures. For Plot, ownership verification, land type (agricultural versus industrial), cable route feasibility, terrain characteristics, and historical risks (unexploded ordnance, archaeology) are assessed. For Profit, revenues are more uncertain than costs; the impact of the curtailment timing on revenue can vary enormously (from 0% to 80% revenue loss at 10% curtailment). Costs are relatively standardised based on experience. Country differences (NL, DE, BE) play a major role across all dimensions.

Additional risks

Trust in and the background of the selling party is an important risk dimension outside the 4P structure. Corporate hygiene of the SPV (annual accounts, KvK registration, directors) is also reviewed. A project may look strong on all 4P dimensions and still be rejected on counterparty grounds.

Requirements for a tool

An overview per 4P dimension with risk flags, severity, probability, and mitigation measures. Traceability is essential: insight into which source documents led to a conclusion. The tool should also explicitly indicate which information is unavailable and which elements are not assessed by the tool. The format should begin with a concise overview of the main findings, with the possibility to drill down into detail. Automation of checks on public sources and comparison with earlier projects are seen as feasible first steps. The tool must be understandable for outsiders (Investment Committee). Permit assessment is considered difficult to automate because of its contextual and relational nature. The tool could partly accelerate and broaden the existing DD process, but not fully replace it.

Interview P2: senior attorney

Date: 15 April 2026 (including post-interview remarks)

Role and perspective

As Senior Attorney, the interviewee is involved in the initial assessment of new project opportunities, primarily from a legal and structural perspective. The interviewee works together with commercial and technical colleagues in a small team that jointly forms a judgement, and emphasises that early-stage project acquisition fundamentally differs from regular company acquisitions: there is still little to investigate, which limits DD to a few core topics. The assessment has two elements: verifying what exists, and estimating what may emerge.

Risk assessment

For the grid connection (Power), the biggest challenge is the unpredictability of the TSO: planning dates are indicative and not enforceable, internal departments within the TSO are not always aligned, and delays occur regularly, as experience within the portfolio illustrates. The interviewee noted afterwards that the accessibility and logistics of a site (access roads for heavy transport) is currently only assessed at a later stage, while this should be mapped earlier. For permits, the assessment in the early stage is deliberately high-level, because many variables are not yet fixed. The interviewee emphasised the importance of identifying early which authority is actually competent for the permit — there has been an instance where municipality and province pointed at each other. It was also noted that BESS projects are new for many municipalities, so civil servants do not always know how to handle them, which causes delays. For the business case (Profit), revenue projections differ significantly between providers, and country-specific market context is critical — the example of Finland, where the energy market structure did not support attractive commercial contracts.

Bottlenecks

Internally, availability of people and funding play a role. Externally, regulatory change is the largest uncertainty. Time is explicitly identified as an underappreciated risk factor: regulatory deadlines, investor exit timelines, and the phasing of investments across multiple projects. There are many assumptions in the early screening process; if DD documentation contains claims (for example shared emails), it would be valuable if these could be automatically cross-verified against external sources.

Requirements for a tool

The tool should function as a completeness check that assesses more data points than are currently covered manually. Source traceability is essential. The tool should also explicitly indicate which information is missing. New requirements raised post-interview include: automated identification of the competent permitting authority based on public documents; retrieval of contextual data on permitting progress in the same municipality or province (for instance benchmarking against comparable municipalities in terms of energy transition activity); automated cross-verification of claims in DD documentation against external sources; and early flagging of logistical risks related to site accessibility. Cross-project and cross-country comparison are considered valuable additions. Detailed scenario analyses are considered unrealistic for the screening phase.

Interview P3: technical consultant

Date: 20 April 2026

Role and context

The interviewee works as a technical consultant within the project development team of the BESS developer. The involvement spans from the first intake until Financial Close for projects offered by prospect developers. A key established point is that no formal screening structure currently exists: projects arrive through diverse channels, and the assessment proceeds via a dynamic group process with senior management. Projects are also rejected for tactical or capacity reasons, independent of their intrinsic quality.

Risk assessment per dimension

A key starting point for the assessment is the information provided by the selling party, typically through an information memorandum, sales brochure, or, in some cases, an initial data room containing documents related to the main project dimensions. This seller-provided material forms the first basis for understanding the project and its apparent maturity. The subsequent assessment then consists of a second validation and completion step, in which additional public information is retrieved to verify, complement and challenge the initial picture before deciding whether to proceed towards an NBO. For Power, connection feasibility is estimated using the TenneT network map, aerial photos of substations (expandability, reserve fields, connected transformers) and public investment plans of the TSO. In foreign markets this is considerably harder due to a lack of data and local knowledge. Three competing TSO priorities determine lead times: renovation of the legacy 1960s–70s transmission network, expansion, and new customer connections. For Plot, public sources (Google Maps, the omgevingsloket, elevation and flood risk maps) are typically sufficient; reachability and heavy transport are flagged as an underappreciated factor. For Permit, procedures differ strongly per country and per Bundesland — the first German permit procedure took approximately one year to understand. For Profit, changing grid tariffs (notably the phase-out of the §118 EnWG exemption in Germany from 2029) and variation between revenue curve providers play a major role.

Bottlenecks and blind spots

The absence of a formal structure means that decisions are partly intuition-driven and hard to compare objectively across projects. Approximately 80% of incoming projects are rejected without the broader team being informed. Country-specific differences in permitting procedures are difficult to estimate; permits are the biggest unknown when entering new markets, while land is the easiest dimension and TSO contact the second-easiest.

Requirements for a tool

The existing 3P checklist is proposed as the foundation, with the tool attempting to retrieve public information automatically per aspect and to display it via traffic-light indicators. Source traceability is essential. The tool must be institutionally embedded and usable by multiple people, not dependent on a single individual. An iterative Scrum-like approach is recommended: start with a prototype within a few weeks, present it to the group, and develop it further. The tool's output should ideally connect to existing reporting flows to the shareholder, and could also better inform the wider organisation about what is in the pipeline. In the screening phase, the specific battery system (supplier, layout, type) is not relevant — what matters is the generic concept fitting the site.

Interview P4: technical consultant

Date: 20 April 2026

Role and approach

As an energy-storage specialist at an independent technical advisory firm, the interviewee's primary activity is conducting independent Technical Due Diligence for BESS projects. The advisor is typically engaged at the moment of a transaction or financing event; scope is adjusted per project phase — early stage focuses on site and grid connection, mid-stage adds permitting, and ready-to-build includes detailed review of contracts and planning.

Risk classification methodology

A structured risk classification methodology is applied: each finding is described as observation and implication, scored on probability and impact (very low to high), and assigned a traffic-light status (red for high risk, amber for medium, yellow for low). Findings are always linked to the reviewed documents with a clear review cut-off date. Where information is absent, this is explicitly flagged as a separate status rather than scored as a risk. All reports pass through a three-layer quality review.

Key risk indicators

The grid connection is the absolute primary showstopper: without it, the project has no basis. The assessment covers both contractual conditions (is the connection bankable, are the conditions firm?) and

timeline realism (is the planned date realistic given the project's actual status?). An HV transformer has a standard lead time of two to three years; if it has not yet been ordered and the developer claims to be ready-to-build in eighteen months, the timeline is unrealistic. For the site, suitability is assessed using public sources (flood risk maps, the cadastre, Google Maps, access routes). Permits are inherently local and uncertain until the moment of formal granting; the zoning plan, the state of communication with the municipality, whether the safety region has been consulted, and noise norms are assessed. Regulatory change is flagged as a risk when there are known discussions (such as on grid tariffs) that may take effect during the project's lifetime.

Bottlenecks

The biggest risk in the early phase is not that something is left entirely unassessed, but that topics fall between the scopes of different advisors — particularly at the boundary between technical and legal, or technical and commercial. Rapidly changing technology and regulation make standardisation difficult; the expert interpretation layer on top of checklists remains essential.

Requirements for a tool

An early screening should focus on the central questions: can this project realistically be built, and by when? A traffic-light system with probability $\times$ impact is the recognised standard methodology. The tool must apply phase-relative assessment (a missing element in the early phase is weighted differently from the same element shortly before financial close), cite source documentation, and explicitly flag absent information. AI support for document summarisation is useful; full substitution of expert judgement is not. An iterative development approach is preferred over a static tool given the pace of market change.

Interview P5: development expert

Date: 22 April 2026

Role and perspective

As a senior development expert, the interviewee oversees projects in Belgium, the Netherlands, and Germany from introduction until the moment the three P's (Plot, Power, Permit) are secured. Both greenfield development and the acquisition of externally developed projects fall within this remit. For acquisitions, a data room is set up containing all available documentation, after which legal, technical, and commercial advisors perform DD work in parallel. A market advisor is considered essential for every project to verify the assumptions in the business model.

Bottlenecks in the current process

Three structural bottlenecks were identified. First, the volume of incoming data exceeds internal processing capacity — documents in the data room are likely to remain internally unread. Second, the quality of internal analyses is undermined by reasoning errors, unverified assumptions, and a lack of source verification. Third, team members each assess their own part without anyone holding a helicopter overview of the total project, leading to missed nuances.

Risk assessment per dimension

For Power, the Belgian connection process is described as a sequence of stages: orientation study, detailed study, realisation agreement offered, realisation agreement signed. The minimum maturity threshold is that the detailed study is at least well advanced; projects with only an orientation study are considered insufficiently mature. For Plot, a clear right to the land is required (option, lease, or purchase contract), but the requirements are less strict than for Power. For Permit, the threshold is more flexible: one project was acquired without a permit application, based solely on a positive feasibility study from a consultancy. An official or semi-official signal of municipal willingness — even short of a formal permit — offers significantly more comfort. A striking difference: in Belgium and Germany there is intensive bilateral contact between developer and authority during the permit assessment; in the Netherlands the process is strictly administrative with no significant interaction. Counterparty risk is mainly mitigated via the payment structure: the selling party receives almost nothing before the 3P milestones are effectively achieved.

Requirements for a tool

The most valuable contribution of a tool would be generating a helicopter overview of the project based on the available documentation, highlighting deviations from the standard. Data room document processing — with automatic extraction of key elements — directly addresses the problem of unread documents. The tool should also test the feasibility of the proposed project schedule against the documented project status. AI is seen as a quality enhancer that enables the BESS developer to ask more precise questions of external advisors, not as a replacement for those advisors. Correctness weighs more heavily than speed. Accountability always remains with a human.

Interview P6: investment expert

Date: 24 April 2026

Role and core principle

P6's role spans all investment activities from acquisition through financing, contracting, and operations. The central principle is that every project evaluation must remain forward-looking, always keeping the end goal in mind — can this asset eventually be constructed and operated profitably? Over-reliance on the prospect developer's own analysis was identified as a structural mistake from earlier transactions.

Regulatory risk as primary showstopper

Regulatory risk — covering both permitting frameworks and grid tariff regimes — is named as the single biggest risk category for any BESS developer. Two types are distinguished: known uncertainty (manageable via scenario stress-testing at 0%, 50%, and 100%) and unknown uncertainty (publicly available information that was not captured by the team). The second category is preventable and is where an AI-supported tool adds the most value — continuously structuring public regulatory information per country, both backward-looking (current decisions) and forward-looking (projected outlook from consultation documents and announcements).

Business case and scorecard

The financial model itself is considered a commodity; what differentiates a good acquisition analysis is whether the underlying assumptions have been validated. A numeric project quality scorecard is called for at acquisition, scoring each 4P dimension on three axes: cost/price, quality, and timing/planning. A project with a signed grid connection agreement scores differently from one where the agreement is due a year from now. Alongside the scorecard, a formal stage-gate process with cross-functional input (technical, commercial, legal) is identified as absent but necessary — the four-eyes principle applied across teams, not just within one.

Requirements for a tool

The single most important reliability criterion for any data-driven tool is up-to-date data. If the regulatory or market data is stale, the tool is not useful. The tool should support assumption validation and provide a forward-looking regulatory context layer per country. It should not build the financial model or claim to eliminate inherent macro risks such as interest rate movements, which are managed through hedging rather than analysis.

Interview P7: engineering consultant

Date: 29 April 2026

Role and involvement

As Engineering Consultant, involvement begins when project development indicates that a site has potential. Engineering then receives a limited dossier and joins the meetings, following the process through to commissioning and operations. The very first screening still happens without engineering; engineering typically becomes involved around the site visit or shortly before.

Technical risk indicators

Noise is the most critical technical risk at site level. The distance to the nearest dwelling is the primary spatial input; the Flemish VLAREM norms prescribe 45 dB(A) at night and 60 dB(A) during the day,

measured at the point of the nearest affected resident. Regional deviations exist (one project had an additional 5 dB restriction imposed by the municipality). Transformer noise is a separate source from battery container noise and must be planned independently in the layout. Container orientation can significantly influence the noise profile; tonality penalties may add a further 3–6 dB to the restriction. Developer-provided noise analyses must be actively challenged on their assumptions.

The cable route is the second-largest technical risk. The distance to the nearest high-voltage substation is the most direct metric; the number of cadastral parcels crossed by the route is a quantifiable indicator of negotiation and legal risk. A minimum trench width of 5 m must be allowed; horizontal directional drilling is more expensive and requires heavier conductors. The connection type and required flex/curtailment percentage is rated a 10/10 risk: high mandatory modulation makes the business case unviable regardless of other factors.

Soil conditions are an underappreciated cost risk, especially for large sites. Relief, bearing capacity, groundwater, and electrical/thermal conductivity determine foundation costs; one past project showed how missing pre-investigation can lead to substantial cost overruns. Pre-acquisition soil investigation should become standard. Accessibility and heavy transport is an additional site-specific factor: bridge weight ratings are the binding constraint for road transport (maximum 40 tonnes in Belgium); waterway access is a strong positive. Seismic sensitivity is a supplementary check that can be performed using public European seismic hazard maps.

Public data sources

All grid operators publish capacity maps; FCR and aFRR volumes and market depth per country are publicly available via TSOs and ENTSO-E (European Network of Transmission System Operators for Electricity); temperature and wind direction data are standard meteorological inputs; cadastre data is public. The interviewee offered to share a personal list of public data sources after the interview.

Requirements for a tool

A geographic, map-based visualisation layer is proposed as the primary interface, overlaying HV lines, substations, cadastral parcels, and congestion status. Concrete functionalities include an automated cable route calculation (minimising the number of cadastral parcels crossed) and a preliminary noise feasibility module based on site location, prevailing wind direction, a reference system specification, and the applicable regulatory limit. Market depth data per country should be retrieved automatically from public TSO/ENTSO-E sources. A traffic-light dashboard with drill-down to geographic detail is the preferred output format. Benchmarking against the BESS developer's existing projects is used to calibrate thresholds. Source attribution and calculation transparency are non-negotiable requirements.

Interview P8: finance expert

Date: 29 April 2026

Role and involvement

As Finance Expert, the interviewee's primary responsibility is structuring project financing once a project has been acquired and is ready-to-build. Due to observed gaps, involvement has recently been extended earlier in the process: reviewing acquisition-phase financial models for errors and for mismatches with the assumptions used later in the project finance phase. This is not a formal task but a necessary addition.

Structural bottlenecks

The greatest bottleneck is that assumptions in the acquisition model are insufficiently substantiated and are not systematically carried forward to the project finance model. Almost all assumptions are generic or based on prior projects, without verification of whether they still apply to the new project. Commercial DD is frequently absent in the acquisition phase, with the consequence that market-specific information — such as TSO operating restrictions or grid fee structures — is not available in time. The German ramp rate example concretely illustrates this: operating restrictions that materially affected the business case were publicly available but were not captured in the acquisition analysis.

Risk indicators and bankability

Regulatory risks regarding grid fees and operating restrictions are the largest threats to equity IRR. As a guideline: an increase in grid fees can move equity IRR from 15% to 7%. For bankability: at financial close, no DD item may carry a “high risk” score; medium risks are acceptable with documented mitigants. Revenue curves are the most uncertain and most impactful input in the acquisition model; in new jurisdictions, the risk is greatest because generic curves can deviate substantially from project-specific curves. CAPEX assumptions are made too simplistically, without accounting for scale differences or site-specific costs. OPEX assumptions are almost entirely ungrounded due to a lack of operations team input.

Requirements for a tool

A standardised list of approximately thirty key inputs per acquisition, with each input tagged with its source, its type (market-generic, region-specific, or project-specific), and its validation status (well-grounded, partially grounded, or assumption). Comparison of the same assumption across multiple projects to identify and explain outliers. Linkage of risk items to their potential impact on the business case. Monitoring of regulatory developments per active market with direct translation into implications. Logging of rejection reasons for projects that did not proceed, so that institutional knowledge is preserved. Source attribution per assumption so that values are traceable and challengeable. Cross-project comparison is the most important adoption driver.

Interview P9: legal consultant

Date: 1 May 2026

Role and perspective

As an external legal consultant at a law firm specialising in renewable energy, the interviewee advises on the legal aspects of project acquisitions in the Netherlands and coordinates transactions for German work, with substantive German legal DD delegated to specialist German counsel. The conversation produced insights that extend beyond the BESS developer, owing to comparative experience from supporting many similar transactions with other developers.

Risk hierarchy and legal certainty threshold

The grid connection is almost without exception the largest risk and the principal reason transactions do not proceed. The legal certainty threshold is clear: only a signed realisation agreement or a signed offer provides enforceable rights — an email from TenneT or a regional grid operator has, at least in the Netherlands and pursuant to a recent injunction ruling, no legal value. TSO timelines are in general systematically missed; a buffer of at least 20% on the date stated by the grid operator is a sensible standard correction. The permit is second in the risk hierarchy: in the Netherlands the municipality is in principle the permitting authority, but the province and waterschap may require additional approvals. Provincial policy must always be independently checked, even when the municipality indicates this is not necessary.

Key aspects of the legal assessment

For each document, the following questions are generally answered from a bankability perspective: is the document revocable or irrevocable? Are there construction or operational restrictions embedded? Are there hidden liabilities or costs? Standard checks also include the chamber of commerce register, the cadastre, internet searches on involved parties, and relevant international sanctions lists.

Structural bottleneck and AI

The most fundamental problem in current DD practice is that advisors work in silos and do not test each other’s work. The financial model is not systematically shared with legal or technical advisors. In the ideal situation, the key line items of the business case model are reconciled with findings from the legal and technical DD reports. AI performs well on the task of checking existing clauses; identifying missing information remains an unresolved challenge, at least as far as renewable energy projects are concerned.

C

Expert Evaluation Sessions

This appendix documents the three practitioner evaluation sessions. Table C.1 gives the overview; the sections below report conduct and substance per session. All sessions were conducted under the consent arrangements of Section 3.6; the evaluative findings are analysed in Chapter 8.

Table C.1: Overview of the practitioner evaluation sessions (2026).

Session	Date	Duration	Participants	Prospect	Goal
Group feedback session	1 July	∼75 min	P1, P3, P5; E1–E3	dummy project (live demo)	Validate the proposed process and the checklist-filling component; collect input for the assessment stage
Record walk-through, business development	7 July	∼60 min	P1	Beta (campaign record)	Audit the extracted content against first-hand project knowledge
Record walk-through, legal / due diligence	27 July	∼105 min	P2	Beta (fresh runs on the frozen version)	Audit content and citation trail from a due-diligence perspective; observe the response to expert feedback

C.1. Group Feedback Session

The session was held with the BESS developer's project development team. Three participants had taken part in the interview programme (P1, P3, P5); P3 is also the author of the manual business-development checklist on which the prototype builds. Three further participants encountered the work for the first time (E1–E3). The mix was intentional: the familiar participants could judge whether the prototype matched what they had described in their interviews, while the newcomers provided a fresh-eyes check on whether the reasoning was followable and whether anything obvious had been overlooked.

Intention of the session

The session had two purposes. The first was to validate the proposed LLM-integrated process, and in particular the checklist-filling component, against the judgement of the people who carry out early-stage screening in practice. The second was to gather structured input on the one part of the process that remained conceptually open, the assessment of the prospect, without allowing the discussion to become an open-ended brainstorm. To keep the session grounded, the researcher set an explicit frame at the outset: the checklist-filling component was presented as the appropriate focus of feedback, while the market-overview and organisational-assessment components were presented as designed on paper only. Participants were asked to give candid opinions rather than encouragement, a request made explicitly to counter the tendency of colleagues to be polite about a thesis.

Conduct of the session

The session opened with a short recapitulation of the current screening process and its bottlenecks, so that the participants who were new to the work shared a common starting point with those who were not. The researcher then presented the proposed four-block process at a high level, before narrowing the focus to the checklist-filling component and demonstrating the working prototype live. The demonstration walked through the creation of a new prospect, the three-phase extraction, and the human-review loop at each phase in which the user confirms, rejects, or flags as uncertain the findings the model has proposed. A dummy project was used for the phases whose live execution would have taken too long in the room.

Rather than issuing a questionnaire, the researcher posed a small number of targeted questions verbally, allowing discussion between them. The questions concerned whether the proposed process was logical and complete, whether the three-way *found* / *not found* / *probable* distinction was useful, what the prototype would change relative to current practice, whether the participants would use the process if it were automated in this way, and whether there was any information they would never trust the model to extract without human confirmation. A final, deliberately broader question invited the participants to consider how the still-open assessment stage should be designed. The discussion was time-boxed by the participants themselves toward the end, and the assessment-stage question was converted into a take-home exercise.

General feedback

The feedback was substantive and, at several points, critically engaged rather than merely affirmative. The main threads are summarised below.

The early, data-poor pathway. P3 repeatedly emphasised that, in practice, much of the value is created before a data room is ever opened, by combining the sparse information in a teaser with public information and domain knowledge, for example inferring a plot's location from knowledge of the high-voltage grid. P5 asked whether the model could assist with exactly this kind of pattern-finding. The researcher agreed but argued that such speculative inference should be kept on a separate track rather than folded into the checklist, so that guessing never contaminates the reliable extracted data. The point was accepted on that basis, and the early teaser-plus-public-sources pathway was noted as a process variant worth sketching explicitly.

Reliability of the public-source step. The public-source phase was identified as the point of greatest hallucination risk. P1 and P3 converged on restricting this step to a curated whitelist of trusted sources rather than an open web search; P3 had already supplied a list of the public sources he uses most, and it was agreed that constraining the model to such sources makes the step markedly more reliable.

Iteration and missing steps. P1 suggested that checklist-filling need not be strictly linear and could instead be iterative, re-running as new documents arrive to fill the list further and to validate earlier assumptions. P3 noted a question-and-answer round with the selling party that the initial process diagram did not show; this was accommodated as a side-process in which the model drafts questions from the gaps identified in the teaser and public sources.

Prioritisation and criticality. P1 observed that the checklist is simultaneously very extensive and not extensive enough, and that it would benefit from some prioritisation of items by their influence. P5 proposed that a defined subset of critical items should always be human-verified even when the model is fully confident, and suggested the team review the checklist line by line to identify them. E1 noted that some items are inherently subjective and cannot be reduced to a binary answer even by a person, which complicates their treatment by the model.

Scope of the model and practical concerns. Several newcomers probed the boundaries of the approach. E2 questioned how the quality of the output would be assured, prompting a discussion of skill refinement as the key remaining optimisation. Questions were also raised about the choice of the underlying software and about token cost. On the former, P5 and P1 clarified that no company-wide tooling decision had yet been made and that the prototype should be understood as tool-agnostic, requiring only an integration with a language model. On the latter, the researcher estimated the cost of analysing a prospect at a few dollars, and at most in the order of tens of dollars for a heavily populated data room.

Positive feedback

Alongside the critical points, the session produced clear affirmation of the work. The transformation of unstructured documentation into structured data was described by P1 as a vital precondition for any subsequent risk assessment. P3 articulated the central benefit in terms of the current workflow, observing that where several people presently read the same documents and record the same findings, the tool can perform that work and leave a single person to validate a shared record for the others. P5 judged that automating the filling of checklists "makes a lot of sense", valued the resulting knowledge base of analysed prospects, and, when asked for an explicit recommendation, endorsed proceeding. The three-way status distinction was called "very useful" by P1. The deliberate narrowing of the thesis onto the checklist-filling component was explicitly supported, with P3 agreeing that the effort should be directed "100%" at that component. The newcomers were able to follow the reasoning and engaged with it in detail, which was itself a useful indication that the proposed process is communicable beyond the group already close to it.

C.2. Record Walk-Through, Business Development

The first walk-through checked the content of the campaign record: P1, who has first-hand working knowledge of Beta (a German prospect), and the researcher went through the filled checklist item by item, comparing the extracted findings, their evidence states, and their cited sources against P1's own knowledge of the project. The conversation was held one-to-one; findings were recorded in the researcher's notes and are reported here and in Section 8.4 in anonymised form, without location, party, or precise capacity details.

P1's overall judgement was positive: for the large majority of reviewed items, the extracted values and their classifications matched P1's knowledge of the project, and P1 considered the record a usable screening basis. One substantive extraction error was identified. On the item recording the seller's motivation to sell (PLOT-17), the extraction attributed a motivation that in fact belonged to a different party in the transaction chain (the landowner rather than the selling entity), and P1 judged that the underlying documents do not support the attribution. The item concerned is flagged in the checklist schema as both subjective and entity-based. The error was corrected through the review interface during the walk-through, exercising the human-gate mechanism (HU-04); its interpretation is given in Section 8.4 and Chapter 9.

Two further observations fed the refinement backlog. First, P1 confirmed the group session's view that items of this subjective, party-related kind should carry a mandatory human check regardless of model confidence. Second, P1 found the per-item source citations the most useful element of the review, since they allowed each doubtful finding to be checked against the named document directly rather than re-searched.

C.3. Record Walk-Through, Legal and Due Diligence

Participant and setting

The second, substantially longer walk-through was held with the senior attorney of the interview programme (P2). P2 was selected for a specific reason: P2 led a large part of the due-diligence process when the BESS developer acquired the prospect used in the session, and therefore holds first-hand knowledge of its non-technical detail (the corporate and deal structure, the land rights, the permit correspondence, and the counterparty relationships). Where the walk-through with P1 tested a business-development reading of the record (Section C.2), this session tested a legal and due-diligence reading of it.

The prospect reviewed was Beta, one of the three deep-run prospects of the demonstration campaign and the data room that had been held back during development (Section 8.1). Its data room holds 232 documents. The session was recorded with P2's knowledge, consistent with the consent arrangements of Section 3.6. It was conducted in Dutch; all quotations below are the researcher's translations, and the original wording of the load-bearing quotations is given in the footnotes. Prospect detail is reported at a level that does not identify the site: parties, location, land areas, and consideration figures are omitted, while document types and checklist item identifiers are retained because the findings concern precisely those.

Method and its deliberate deviation

The session was designed as a naturalistic, formative, ex-post evaluation in the terms of Venable, Pries-Heje, and Baskerville [98]: a real prototype, a real data room, a real expert, and a real prospect, but not a live investment decision. The format was a concurrent verbalisation over the prototype's output, in the tradition of verbal-protocol analysis [28]: the expert inspected each item in the review interface and articulated a judgement on the finding, its reasoning, and its cited source, while the researcher recorded the verdict and the reason for it in the tool.

Two deviations from a stricter protocol should be stated plainly, because they bound what the session can establish. The first is the order of elicitation. A blind protocol would have asked P2 for an independent answer before revealing the model's finding. P2 had not visited this data room for a considerable period, and cold recall across 152 items would have produced mostly non-answers. The extraction was therefore shown first, together with its evidence state, its reasoning, and its cited quote. The consequence is that this session measures an expert audit of the record against its own citation trail, not an independent reconstruction of the record from memory. Agreement figures from a protocol of this kind are upper bounds.

The second is the researcher's participation. The researcher was not a silent observer: several of the interpretations reported below were put to P2 by the researcher and then accepted, refined, or rejected. This is a recognised position in interpretive field research rather than a defect, provided it is disclosed [100]. Three features of the session limit the risk that P2's judgements were displaced by the researcher's. The verdicts were typed into the tool as P2 spoke, in P2's own words, and the resulting text is preserved literally in the prompt trace. The dominant finding of the session was volunteered by P2 rather than proposed by the researcher. And P2 declined to answer, or rejected the premise of, a number of items instead of accommodating them. Where a position below was reached jointly, it is reported as such.

Conduct

The session followed the three phases of the pipeline in order. It began with the quick phase, run from the teaser material alone, so that P2 could see what the record contains before a data room exists. Each of the 52 in-scope items was reviewed; P2 confirmed, corrected, or supplied values, and the researcher recorded the edits. The phase was then declined with P2's inline corrections, re-run, and the changed record reviewed again.

The full phase was then executed against the data room. During the run, which took approximately nine minutes, the token and cost accounting was shown; P2 asked how the spend was tracked and where it was attributed, and the observability of the cost ledger was itself noted as a useful property. The 35 probable findings and the critical values were then reviewed item by item, followed by a walk through the remaining dimensions. Where P2 judged a finding unsupported, the researcher recorded a per-item rejection reason. The phase was declined with seventeen such rejections and a general instruction, re-run overnight, and the second version audited afterwards outside the tool.

The session closed with four open questions: what the largest remaining development points are, what P2 found positively surprising, what share of the answers P2 considered genuinely correct, and where the boundary of trust should sit.

Runs performed

All runs used the frozen evaluation version of the prototype. The runs themselves are new, executed on 27–28 July 2026 on a fresh instance of the prospect; their counts therefore differ from Beta's campaign row in Table 8.2, which reports the runs of 7 July. In this instance, the file-type filter skipped fifteen documents in unsupported formats.

The quick-phase re-run was marginally more conservative: two items moved from found to not found. The full-phase re-run raised the found count by a third and converted a net eight probable findings into confirmed ones. Whether it acted on the substance of the feedback is a separate question, treated below.

Table C.2: Runs executed for the walk-through, on the frozen prototype.

Run	Model	Duration	Found	Probable	Not found	Out of scope
Quick v1	Haiku 4.5	38 s	22	0	30	100
Quick v2	Haiku 4.5	42 s	20	0	32	100
Full v1	Sonnet 4.6	8.9 min	30	35	57	30
Full v2	Sonnet 4.6	6.5 min	40	27	55	30

The dominant finding: source substitution

P2's central objection was not that the extraction was wrong. It was that the extraction was, repeatedly, right about the wrong document. Across the full phase, the model preferred a single seller-authored timeline overview to the underlying documents from which that overview was itself compiled.

P2's reasoning for why this matters is a due-diligence argument rather than a preference. Due diligence exists to establish whether the seller's account holds: "Due diligence is there to validate whether what the seller says is correct. What he says is in his information papers. Whether it is true follows from the underlying documents."[1] A record that cites the seller's own summary therefore does not support the task the record exists to serve. Asked what he wanted instead, P2 was concise: "The source. Not a sales brochure based on the source, but the source itself."[2]

P2 was explicit that a seller statement is not worthless, only insufficient: "A seller's statement that something is the case is worth something to me, but I would rather be shown the document it follows from." A second objection concerned what such overview documents actually contain. They record the existence of a study or a milestone, not its content: "It is only mentioned as an item. It does not say what there is; it says that there is something."[3] This explains a pattern visible throughout the record, in which the reasoning field concedes that no specific information was available. P2 identified the cause: the information exists, but not in the document the finding cites. P2 characterised the behaviour as "taking the shortcut to a small document where the matter is mentioned".

The clearest instance concerned a critical value. After the full run, the record still carried the lower capacity figure taken from the seller's timeline overview, whereas the transaction had proceeded on a higher figure. P2 stated where the authoritative evidence would be, and the two letters from the transmission system operator confirming the higher capacity were then located in the data room during the session. P2's comment on finding them was that these are the sources the value actually follows from, and that the record ought to say so. The consequence is not confined to one item: a derived price-per-megawatt figure elsewhere in the record had been computed against the superseded capacity.

Composition of the cited sources

To test the objection systematically, the citations of the second full run were audited manually against the type of document cited and whether that document was composed by the selling party. Before the audit figures can be read, the funnel that determines what the model sees must be stated. The data room holds 232 documents. The file-type filter removed 15 documents in unsupported formats (among them the seven binary spreadsheets holding the question-and-answer record); the triage step tagged all 217 remaining candidates as medium relevance, providing the budget no discrimination; and the page budget then selected 44 documents across the run's requests. Table C.3 reports the audit.

Two figures carry the finding. Eleven documents were cited, which is 4.7% of the room and 25% of the 44 documents the model actually received. And 61.3% of all quotations came from documents the selling party had itself composed, with a single timeline overview supplying more than a third of the total.

Asked what these figures should look like for a record he would consider adequate, P2 gave two benchmarks, both volunteered in response to the audit rather than proposed by the researcher. Table C.4 sets them against the observed values. The two columns are not directly commensurable: the routing

[1] "Want die DD is om te valideren of wat die verkoper zegt klopt. Dus wat hij zegt, staat in zijn informatiepapieren. Maar of dat klopt, dat blijkt uit …de onderliggende stukken daarvoor."

[2] "Ja, de bron. Niet zeg maar een verkoopbrochure op basis van de bron, maar de bron zelf."

[3] "Het wordt alleen genoemd als item. Er staat niet wat er is, er staat dat het er is."

Table C.3: Composition of the cited sources in the second full run ($n = 75$ quotes).

Type of source	Composed by seller	Share of quotes
Timeline overview	yes	34.7%
Teaser: prospect specifics	yes	14.7%
Technical drawing, substation connection	no	10.7%
Teaser: deal structure	yes	6.7%
Commercial register, seller	no	6.7%
User edit (no specific document)	n.a.	6.7%
E-mail chain, teaser discussion	no	5.3%
E-mail chain, Q&A clarifications	yes	5.3%
Proof of payment, bank	no	4.0%
E-mail, non-binding offer	no	1.3%
E-mail, revised purchase offer	no	1.3%
Draft share purchase agreement	no	1.3%
Quotation without specific document	n.a.	1.3%
Composed by the selling party		61.3%
Independent of the selling party		30.7%
No specific document		8.0%
Distinct documents cited		11
Share of the data room cited (of 232)		4.7%
Share of the routed documents cited (of 44)		25%

budget caps how many documents reach the model at all, so the observed coverage is bounded below the benchmark by construction. The comparison locates the distance to practice; it does not attribute all of that distance to the model.

Table C.4: Observed citation properties against the practitioner benchmarks.

Property of the record	Observed	Benchmark (P2)
Share of data-room documents cited	4.7%	40–50%
Share of quotations from seller-composed documents	61.3%	5–10%
Largest single source as a share of quotations	34.7%	—
Distinct documents cited	11	—

A separate, mechanical check applies to the same run. The tool verifies every quoted citation against the extracted text of the document it names. Of the citations it could check, roughly a quarter could not be located in the cited document at all.[4] P2 encountered this class directly on several items, where the value stated was correct but could not be traced to the document named.

Whether the re-run acted on the feedback

The seventeen rejected items and the accompanying general instruction were fed into the second full run. Three items moved to a different and more specific source. Fourteen continued to cite the same seller-authored timeline overview that the feedback had asked the model to move away from. The three that moved are exactly the three whose rejection reasons contained a concrete factual correction. The fourteen that did not move are those whose rejection reason was the general instruction to prefer more specific, non-seller sources. On the evidence of this single round, the model responded to specific corrections and not to a general sourcing principle.

Two mechanisms could produce this, and the session cannot separate them. The instruction may have been too abstract relative to the item lines, which continue to direct the model where they previously directed it. Or the underlying documents may not have reached the model at all: the routing step selects a subset of a large room under a page budget, and during the session the permit correspondence had

[4] The manual audit and the tool's automatic count do not reconcile ($n = 75$ against $n = 48$ citations, and eleven distinct documents against four). The two were produced by different procedures and the discrepancy is unresolved; the manual audit is used for the composition analysis and the machine check only for the order of magnitude of unlocatable quotations.

to be opened manually in the data room before it could be seen. Both mechanisms are consistent with the observation that the authoritative capacity letters existed in the room and were not cited. P2 proposed the experiment that would discriminate between them: withhold the overview documents and re-run. “If you take the overviews out, it cannot give them as a source. Then it has to go to the real document.”[5] This test was not performed within the scope of this thesis and is recorded in Chapter 9 as the immediate next step.

Item specification as the second development point

Asked at the close what the largest remaining development points are, P2 named the specification of the checklist items before anything else: the questions themselves need to be tighter, and doing so “would solve a great deal”. Earlier in the session P2 had put the point more sharply: “If you regard the left-hand column as a prompt, then this is a bad prompt.”[6]

The observation is structural rather than incidental. In the implemented design each checklist item contributes a generated line to the extraction prompt, so an under-specified item is an under-specified instruction. P2 reached this by working through items whose text admits several readings, and the pattern recurred across dimensions. Table C.5 records the instances discussed and the reformulations agreed in the session.

Table C.5: Item-specification problems identified in the walk-through.

Item	Problem identified	Position reached
SELL-02 company structure	Answered with the identity of the seller, duplicating an earlier item.	Should record the object of the transaction, including subsidiaries and entities.
SELL-04 local country residence	Does not state whose address is meant; seller’s registered address and project location are different things.	Specify the subject of the address explicitly.
PLOT-11 seller name and address	Under a plot heading, was answered with a share-transaction seller rather than the landowner.	Under Plot the question can only concern the party that sells or leases the land.
PLOT-08 expansion options	The documents give a current area and a maximum; the item asks for the additional area. The model reported the raw phrase.	The additional area is derivable and should be the recorded answer; the maximum is ambiguous between aggregate and additional.
PLOT-10 distance to substation	Framed as a threshold question, so a precise known value was replaced by the coarser threshold answer.	Where the precise value is known it should be recorded.
POWER-12 distance to power line	P2 had not encountered the concept in documents of this kind and questioned its relevance.	Clarify whether a substation connection or a pylon tap is meant, or reconsider the item.
PLOT-40 third-party rights	Too broad to answer: rights of whom, over what.	Narrow to named right-holder classes.
PERMIT-20 prof. stakeholders	P2 declined the item as unanswerable as written.	Enumerate the stakeholder classes intended.
PERMIT-22 preliminary design	Answered with permit-regime material rather than the state of the site design.	Separate design readiness from planning regime.
PERMIT-09 groundwater	“Water” may mean groundwater, surface water, or drainage.	Split into the distinct questions.
PERMIT-16, PERMIT-03	Dutch statutory terms carried onto a German prospect, where no corresponding instrument exists under that name.	Ask the question, e.g. whether an environmental impact assessment is required, and resolve the jurisdiction-specific instrument per country.

The jurisdiction problem was the one P2 pressed hardest. Several permit items are named after Dutch statutory instruments. On a German prospect the model cannot find them, because the instrument does not exist under that name, and a not-found verdict on such an item carries no information. P2’s

[5] “Als je die eruit haalt, dan kan hij dat niet als bron geven. Dan moet hij naar het echte stuk gaan.”

[6] “Maar dit is zeg maar een beetje, als je de linkerkolom als een prompt ziet, dan is dit een slechte prompt.”

proposed remedy was to pose the functional question and let the jurisdiction-specific instrument be resolved separately. The researcher noted that this is the role the market-overview block of Section 7.1 is intended to play, maintained per country and fed into the extraction.

Errors observed, by class

Beyond source substitution, the walk-through surfaced several distinct failure modes. They are grouped in Table C.6 because the grouping, rather than any single instance, is what the session establishes.

Table C.6: Classes of error observed in the walk-through.

Class	Observation	Design implication
Source substitution	A seller-authored overview cited where the underlying document exists in the room. Fourteen items after re-run.	Document-set composition and source-preference rules, not prompt wording alone.
Unsupported citation	Value correct, cited document does not contain it.	Provenance is only as good as its weakest citation; the mechanical quote check should gate the finding, not annotate it.
Reasoning contradicts finding	The reasoning field records that a matter is undetermined; the finding asserts a determinate value.	The two fields should be checked against each other before the record is presented.
Precision regression	A precise value established at an earlier gate replaced by a coarser threshold answer.	Confirmed values must constrain later phases.
Stale value on a critical item	Superseded headline capacity retained, propagating into a derived figure.	As above; and derived figures must be recomputed when an anchor changes.
Entity conflation	Landowner, selling entity, and share-transaction seller interchanged; a cable rating reported as a substation voltage.	The error family already seen with P1; entity-typed items need explicit subjects.
Hedged answer on a binary item	A payment reported as made or in progress, where the item requires one or the other.	“One answer is yes, the other answer is no”; hedging is not the same as declared uncertainty.

Two of these have a common and identifiable cause in the architecture. Human-confirmed critical values are injected into the prompt of the gaps phase but not of the full phase (Appendix D.2). The capacity correction P2 made at the quick gate was therefore invisible to the full extraction, which re-derived the value from the seller’s overview; and the precise distance confirmed at the quick gate was likewise unavailable when the full phase answered the same item as a threshold. Both regressions are consequences of where the anchor mechanism is wired, not of model behaviour, and both are correctable.

Positive findings

P2’s positive observations were specific rather than general, and are reported as given. The clearest concerned robustness to weak specification. What P2 found positively surprising was that the model still retrieves the most obvious relevant information even when the question is unclear: on a scale of warm, warmer, warmest, “it still gets to warm”.[7] P2 added that where an item admits several interpretations, the model tends to arrive somewhere defensible.

Asked to estimate the share of genuinely correct answers, P2 put roughly three quarters of the probable findings in that category, describing the estimate as off the cuff, and observed that the model performs comparatively well precisely where it declares its own doubt. The estimate is bounded in two ways that must travel with it: the review concentrated on the items the model was uncertain about, and it is a single expert’s impression rather than a measurement. In the same breath P2 noted that the model “sometimes misses nuances” and that “there are sometimes sloppy bits in it”.

[7] “Dus dat hij wel, zeg maar, als je warm, warmer, warmst hebt, dat hij nog wel warm is.”

The reasoning field was found to carry genuine value independently of the finding. On the item where a cable rating had been reported as a substation voltage, P2 accepted the reasoning as correct on the ground that it recorded the matter as unclear, and the item was accepted on that basis. The per-item citation and the click-through to the cited page were used continuously throughout the session as the working mechanism of the review, which corroborates the same observation from the P1 walk-through.

Several substantive findings were confirmed against P2's own knowledge, including the non-disclosure arrangement and its date, the seller's intention to remain involved, the readiness of the data room for due diligence, and the deal structure as proposed. On the last of these P2 noted that the structure recorded is the seller's proposal, that the BESS developer subsequently negotiated a different one, and that this later change is not in the data room and so could not have been extracted.

Observations on the evidence model and the record

Three refinements to the record itself emerged. The first is a missing state. P2 distinguished a blank answer from an answer of not applicable, and treated the difference as material to any due-diligence report: a blank leaves the reader unable to tell whether the matter was examined, whereas an explicit not applicable establishes that it was examined and found absent or irrelevant. The three-state evidence model of Section 7.2 does not currently carry this distinction. The point converges with E1's observation in the group session that some items cannot be reduced to a clean binary answer.

The second concerns conflicts. The conflict view compares values textually, so semantically identical values are surfaced as conflicts, and several of the conflicts reviewed were of this kind. More substantially, the interface requires the reviewer to choose one of two values, whereas on at least two items P2 judged that the correct answer was the combination of both and that neither alone was complete. A merge action, and a semantic-equivalence filter ahead of the conflict list, both follow from this.

The third concerns items that documents cannot answer. Some items describe prototypes the BESS developer holds rather than the seller, and cannot be extracted from a seller's data room at all. Others record activity that leaves no document: on contacts with local authorities P2 observed that these may amount to many conversations and meetings of which nothing is minuted. Such items are not extraction targets, and treating a not-found verdict on them as an extraction failure misreads the record.

Overall judgement

The session's overall verdict is that the process is on the right track but that the instrument is not yet sufficiently specific or thorough for the record to be relied on in due diligence. The two development points P2 identified as decisive were, in his own summary, the left-hand column and the requirement that the model go to the source rather than to the seller's own overviews. Neither is a question of model capability. Both concern the specification of the instrument and the composition of the documents it is given, which are properties of the design.

D

Prototype Transparency

This appendix documents the prototype of Chapter 7 at the level of detail that reproducibility requires and that the chapter itself cannot carry. Five kinds of material are collected here. Section D.1 specifies the checklist instrument that the prototype fills, together with the rules by which the findings of successive phases merge into one record. Section D.2 describes how a request is assembled and reproduces the complete instruction files that drive the model. Section D.3 specifies what is sent to the API per call type, what is recorded of every call, and what a run costs. Section D.4 illustrates the tool from the reviewer's perspective on a fictional prospect. Section D.5 condenses the development history that produced this version.

Everything documented here is the frozen evaluation version. Its instruction texts, checklist version, and model settings were fixed on 7 July 2026, and the demonstration campaign of Chapter 8 was run entirely on that version. Where a file is reproduced, it is reproduced literally, including the notes addressed to the developer rather than to the model, with one exception: example file names inside the instruction texts that referred to a real prospect are replaced by their fictional (Meridian) equivalents.

Two boundary statements frame the appendix. First, the instruction files of Section D.2 are the only prompt sources of the extraction calls; code comments or docstrings are never sent to the model. The one auxiliary exception is the data-room triage call (Table D.5), whose short classification instruction is assembled by the routing code and contains no extraction instructions, only the file listing it classifies. Second, the model receives the prospect's documents as original files (native PDF, with spreadsheets and e-mails converted to text at upload). It never receives the researcher's notes, the thesis, or material belonging to another prospect.

D.1. The Checklist Instrument

This section documents the checklist instrument (schema version 6) that the prototype fills. The instrument formalises an internal business-development checklist of the BESS developer into a machine-readable schema of 152 items. Below, the schema fields are defined first, the descriptive statistics of the instrument second, and a representative excerpt third. The section closes with the rules by which the findings of the three phases merge into the master record.

Schema

Each checklist item is a record with 12 fields, listed in Table D.1. The three phase flags share one value set: `y` (expected in this phase), `m` (may be present; an absence is acceptable), `c` (cross-check: answered from documents earlier, verified against public sources in the gaps phase), and `n` (out of scope for this phase).

Descriptive statistics

Table D.2 gives the composition of the instrument, by factor and by the properties that determine how each item is treated.

Table D.1: Fields of one checklist item (schema version 6).

Field	Meaning
`item_id`	Stable identifier, prefixed by factor (e.g. SELL-01, POWER-03)
`factor`	One of: Selling party, Plot, Power, Permit, Profit
`dimension`	Short item title
`what_to_extract`	Precise extraction definition sent to the model
`specific_public_source`	Suggested public source(s), per jurisdiction where relevant
`for_quick_data`	Phase flag: teaser-based quick extraction
`for_full_data`	Phase flag: data-room extraction
`for_internet_extr`	Phase flag: public-source fill or cross-check
`manual_confirmation`	`y` marks a critical value a human must confirm
`subjective`	`y`/`m` marks items without a clean binary answer
`location_based`	Item depends on the site location
`entity_based`	Item depends on the identity of a party

Table D.2: Composition of the instrument (152 items).

Factor	Items	Property	Items
Plot	71	Quick scope (`y`: 5, `m`: 47)	52
Power	26	Full scope (`y`: 39, `m`: 83)	122
Permit	22	Gaps scope (`y`: 54, `m`: 35, `c`: 25)	114
Selling party	19	Critical (`manual_confirmation` = `y`)	11
Profit	14	Subjective (`y`: 8, `m`: 9)	17
		Location-based	108
		Entity-based	26

The composition matters for the design in two ways. The dominance of Plot items (71 of 152) reflects the observation of Chapter 5 that plot- and permit-related information decomposes into many concrete, checkable sub-aspects at the screening stage. The high share of location-based items (108) is what makes public-source verification feasible: cadastres, grid maps, and zoning portals answer to a location.

Representative excerpt

Table D.3 reproduces eight items from the schema, chosen to span the five factors and the flag combinations that drive the pipeline (a critical anchor, a cross-check item, a subjective item, and jurisdiction-specific public sources).

Table D.3: Extract of eight items of the checklist instrument.

ID	Dimension	What to extract	Public source	q/f/i	crit	subj
SELL-01	Name	Legal entity name of the ACTUAL selling party (current owner of the project/SPV), not the sell-side adviser running the sale (recorded under SELL-18)	Company website; KVK (NL) / KBO-CBE (BE) / Handelsregister (DE)	y/y/c	y	n
SELL-18	Sell-side adviser	Legal entity name of the adviser/bank running the sale process; ‘none’ if the owner sells directly	Registries	m/y/c	n	n
POWER-01	Size (MW)	Plant output/input power (MW)	–	y/y/n	y	n
POWER-03	Storage capacity (MWh)	Capacity	–	y/y/n	y	n
PLOT-03	Plot coordinates	WGS84 decimal latitude, longitude of one representative point, formatted exactly ‘lat, lon’; conversions from local grids go in reasoning	Kadaster (NL) / FOD Financiën (BE) / Bundesland cadastre (DE)	m/y/c	y	n
PLOT-17	Seller reasons to sell	Motivation	Public records; press	m/m/y	n	y
PERMIT-03	Actual zoning plan	Current designation	omgevingsloket.nl (NL) / geopunt.be (BE-Fl)	n/m/y	n	n
PROFIT-01	Grid fee exemption	Eligibility	Regulator: ACM (NL) / CREG (BE) / BNetzA (DE)	m/m/m	n	m

The complete 152-item instrument derives from internal company material and is therefore not repro-

duced in full in this public document; it is available in the confidential digital appendix accompanying this thesis. The excerpt above, together with the schema definition and the statistics, suffices to understand and reconstruct how the pipeline is scoped and driven.

Merge and trust rules of the master record

Chapter 7 summarises the trust hierarchy by which phase results merge into the master record. Table D.4 states the complete rule set as implemented in the engine.

Table D.4: Merge and trust rules of the master record.

#	Rule
1	Every phase's finding is appended to the item's per-phase history; snapshots are immutable audit records.
2	Disagreement with the current value creates a conflict with full records of both sides; quantitative tolerance is 1% (critical) / 5% (other); blocking conflicts gate approval.
3	User ownership is sticky: an edited, confirmed, or ruled-on item is never silently overwritten by a later phase.
4	A phase that finds nothing never erases an earlier finding.
5	Gaps-phase cross-checks never replace a document value; they confirm it, fill a hole, or raise a conflict.
6	A stale critical confirmation reopens when a later, better-informed phase finds new or different data.
7	A rejected finding travels: on decline and re-run, the rejection reason goes to the model as targeted feedback, and the fresh finding is judged on its own merits.

D.2. The Skill Files

This section makes the prototype's interaction with the language model inspectable. It first describes how one request is assembled, and then reproduces every instruction file of the frozen version literally. The term *skill* is used in the sense of Section 3.5.1: a documented prompt artefact with a specified task, defined inputs, and an explicit instruction structure. The API is stateless, so a single request is the model's entire context, and everything the model knows about a prospect is assembled from the parts described below.

Anatomy of one extraction request

Every extraction request is assembled from four parts.

1. **Document blocks.** The prospect's files, each titled with its exact relative path so that the model cites real filenames. The final block is marked for prompt caching.
2. **One instruction block.** The shared general instructions, a separator, and the phase file with its placeholders substituted. Three placeholders exist. `{{items}}` becomes one generated line per in-scope checklist item, of the form `item_id | factor / topic | extract: … | look at: …`, with cross-check lines in the gaps phase additionally carrying the prior document finding to verify. `{{anchors}}` becomes the confirmed critical values. `{{feedback}}` becomes the reviewer's accumulated decline feedback and per-item rejection reasons, or an explicit first-attempt note.
3. **A forced tool schema.** The model must answer by calling `record_findings`. Per item, the schema demands the evidence state (found, probably_found, not_found), the bare finding, the reasoning, the evidence source, the exact evidence document and 1-based page, a literally quote, and a confidence level. Forcing the tool makes every response machine-parseable by construction.
4. **Generation settings.** The model and the output-token cap are read from the phase file's front matter; temperature is fixed at zero.

The five files below are reproduced in the order in which the pipeline uses them: the shared instructions that open every request, the site-detection call, and then the three phase files.

Shared instructions (all phases)

File `general_instructions.md`, sent first in every request of every phase, before the phase-specific file.

```
# General instructions (all checklist phases)
## The documents

The prospect's documents (the PDF teaser and any dataroom files) are provided to
you **as the original files** - not as extracted text. Read the tables, maps,
site drawings, charts and photos inside them exactly as a human reviewer would;
never assume information is unavailable just because it only appears in an
image or figure.

## Who Claude is

You are a due-diligence analyst for Battery Energy Storage System (BESS) project
acquisitions. Your job in this step is **retrieval only**: you read the uploaded
documents and record what they say. You never score, rate, or give an opinion on
whether the prospect is a good deal - a separate step does that later.

## The three evidence states

For every checklist item you must choose exactly one state:

- **found** - the value is clearly stated (or clearly shown in a table/figure) in
  the documents. A source is required.
- **probably_found** - the documents strongly imply the value but do not state it
  outright. Put your best reading in the finding and explain the doubt.
- **not_found** - the documents do not contain the information. Leave the finding
  empty.

Prefer **not_found** over guessing. A confident "not found" is more useful than an
invented value.

## Confidence

Separately from the state, rate your confidence as **low**, **medium**, or
**high**. Confidence is about how sure you are, independent of the state - e.g.
"not_found" + "high" means you are sure the information is genuinely absent.

## Finding vs. reasoning - keep them separate

Every item has two separate fields:

- **finding** - the value itself, and nothing else. For a quantitative item
  (a power, a capacity, a distance, a count, a price) the finding must be **just
  the number with its unit** - e.g. `350 MW`, `1400 MWh`, `<200 m`. Do not put
  sentences, caveats or derivations in the finding.
- **reasoning** - where the derivation, assumption or caveat goes. If you
  computed or inferred the value, explain it here - e.g. `implied by 4-hour+
  duration at 350 MW`. If the value is stated outright and needs no explanation,
  leave reasoning empty.

Example: the teaser says "350 MW, 4-hour duration". You record
finding = `1400 MWh`, reasoning = `derived from 350 MW × 4 h duration`.

**Multi-site / portfolio teasers:** when a project-level item (like total MW or
total MWh) covers several sites, the finding is still just the total number -
e.g. finding = `850 MW`, reasoning = `total portfolio: Project I 350 MW, Project
II 500 MW`. Never write the per-site breakdown into the finding itself.

## Language and number formats

- The documents are often in German or Dutch. Always record findings, reasoning
  and source descriptions **in English** - but keep verbatim quotes in their
  original language.
- Normalise numbers to international notation: decimal **point**, no thousands
  separators in the finding (German `1.400,5 MWh` → finding `1400.5 MWh`).
  Legal/register identifiers (e.g. `HRB 140978 B`) stay exactly as written.
- Dates: use an unambiguous form - `15 Mar 2027`, `Q3 2026`, or `2026` as
  precise as the document supports; never numeric day/month forms like
```

```
  `03/04/2026`.

## Rules that always apply

- Never invent a value, a source, or a quote. If it isn't in the documents, it is
  not_found.
- Copy quotes **verbatim, character for character** from the document - quotes
  are machine-checked against the cited file, and a paraphrased "quote" will be
  flagged to the reviewer as unverifiable.
- If two documents contradict each other, record the value from the most recent
  or most specific document, and state the contradiction in reasoning (name both
  documents and both values) - contradictions inside a dataroom are themselves
  due-diligence signals.
- When a value **is** found, give the source and, where the evidence is text, a
  short verbatim quote. Cite the source as the document's **exact title/filename**
  plus the page or section - e.g. `Meridian milestone structure.pdf, page 3` or
  `Meridian_Project_Overview.xlsx, sheet 'Projects Overview'`. Never write
  "Document 1" or "the teaser" - the reviewer must be able to open the named
  file directly.
- Additionally fill the structured citation fields: `evidence_document` = the
  exact filename (as titled), and `evidence_page` = the 1-based page number when
  the document is a PDF. These power a click-to-open link in the review UI.
  Leave them empty for not_found items and website sources.
- Keep the finding short and factual - the value, not a paragraph. Put any
  explanation in reasoning, per the section above.
- In the finding, reasoning, source and quote text, use single quotes (') instead
  of double quotes (") so the data stays clean.
- Only work on the items you are given in the phase-specific list below. Never
  touch items that are not in that list.
```

Site detection

File `site_detection.md`, the separate call that opens the quick phase.

```
---
title: Site detection
phase_key: site_detection
model: claude-haiku-4-5
max_tokens: 1500
---

# Site detection

<!-- Editor note: runs once at the start of the quick-extraction phase, as a
separate call with its own tool (record_sites). general_instructions.md is sent
first, then this text. -->

Read the uploaded teaser and list the distinct BESS sites or sub-projects it
describes. Many teasers are a single site - then return exactly one entry. A
portfolio teaser lists several (e.g. 'Project I', 'Project II') - return one entry
per site with that site's own MW / MWh / development stage where stated.

Do not invent sites; only split into multiple entries when the teaser clearly
describes separate sites at **different locations**. Construction phases or
expansion stages of the same site (e.g. 'Phase 1: 100 MW, Phase 2: +150 MW' on
one plot) are ONE site, not several. When in doubt, return a single entry.

## Output

Call the `record_sites` tool with the result (one entry per site).
```

Quick extraction phase

File `quick_extraction.md`, the teaser-based phase.

```
---
title: Quick extraction
phase_key: quick
phase_flag: for_quick_data
model: claude-haiku-4-5
max_tokens: 25000
```

```
---

# Quick extraction phase

**Input:** the teaser / pitch documents.
**Goal:** fill only the small set of headline items that a teaser normally
contains, so the prospect can be summarised at a glance.

## Feedback from the human reviewer

If this is a re-run, the reviewer declined the previous attempt(s) with the
feedback below. Address it explicitly - where it conflicts with a default
instruction, the reviewer's feedback wins:

{{feedback}}

## Which items you fill in this phase

You are given the list of items whose `for_quick_data` flag is **y** (expected to
be present in a teaser) or **m** (may be present; not_found is perfectly
acceptable). Do not attempt any other item.

## What to focus on

The teaser almost always states the headline power figures, the seller's name and
the development stage. Pay special attention to these anchor values, because the
later phases build on them:

- Selling party name (SELL-01)
- Size in MW (POWER-01) and storage capacity in MWh (POWER-03)
- Voltage level (POWER-02) and the responsible TSO/DSO (POWER-04)

For everything else in the list, record it if the teaser states it and mark it
not_found otherwise. A teaser is short - some `m` items will be not_found, and
that is fine.

## Items to fill in this phase

Fill each checklist item listed below - one finding per item, and only for these
items. Each line is `item_id | factor / topic | extract: <what to extract> | look
at: <public source>`.

{{items}}

## Output

Call the `record_findings` tool with one entry per item you were given.
```

Full extraction phase

File `full_extraction.md`, the data-room phase.

```
---
title: Full extraction
phase_key: full
phase_flag: for_full_data
model: claude-sonnet-4-6
max_tokens: 50000
---

# Full extraction phase

**Input:** the documents attached to this request. For small prospects that is
the full dataroom; for large datarooms it is a routed selection of the documents
most relevant to the items below (chosen by a triage step).
**Goal:** fill the detailed checklist from the attached documents.

## Feedback from the human reviewer

If this is a re-run, the reviewer declined the previous attempt(s) with the
feedback below. Address it explicitly - where it conflicts with a default
```

```
instruction, the reviewer's feedback wins:

{{feedback}}

## Which items you fill in this phase

You are given the list of items whose `for_full_data` flag is **y** (high
expectation the dataroom contains it) or **m** (medium expectation; not_found is
acceptable). Do not attempt any other item.

## How to work

- Treat every uploaded document as the dataroom. Read tables, appendices, site
  drawings and figures - not just the running text.
- `y` items should usually be answerable from a complete dataroom; if a `y` item
  is genuinely absent, mark it not_found with high confidence so the gap is
  visible.
- For each item, set the state (found / probably_found / not_found), give the
  source and a short quote when found, and rate your confidence.
- Keep the anchor values (seller name, MW, MWh, voltage, TSO/DSO, deal structure)
  precise - downstream steps depend on them.

## Items to fill in this phase

Fill each checklist item listed below - one finding per item, and only for these
items. Each line is `item_id | factor / topic | extract: <what to extract> | look
at: <public source>`.

{{items}}

## Output

Call the `record_findings` tool with one entry per item you were given.
```

Fill-the-gaps phase

File `fill_the_gaps.md`, the public-source phase. The note at the end of the file is addressed to the developer and documents the structural cost controls on live web search.

```
---
title: Fill the gaps
phase_key: gaps
phase_flag: for_internet_extr
model: claude-sonnet-4-6
max_tokens: 32000
web_search: on
---

# Fill the gaps phase

**Input:** the attached documents (for large datarooms: a routed selection
relevant to the items below), the confirmed anchor values, plus the public
sources listed per item.
**Goal:** fill the items that are normally answered from public / internet
sources rather than from the dataroom, and cross-check the items that were
already answered from the dataroom against those public sources.

## Confirmed anchor values

These are the values already confirmed for this prospect. Use them to identify
*which* company, site and project to look up when you consult a public source
(e.g. search for this exact selling party, at this exact location):

{{anchors}}

## Feedback from the human reviewer

If this is a re-run, the reviewer declined the previous attempt(s) with the
feedback below. Address it explicitly - where it conflicts with a default
instruction, the reviewer's feedback wins:

```

```
{{feedback}}

## Which items you fill in this phase

You are given the items whose `for_internet_extr` flag is set. Each item line
tells you what to extract and which `specific_public_source` to look at first
(a cadastre, a grid map, a regulator, google maps, a company register, etc.).

There are two kinds of line:

- **Fill items** - no prior finding is shown. Extract the value from the named
  public source (or the documents if they already contain it).
- **CROSS-CHECK items** - the line ends with `CROSS-CHECK against public
  sources - prior document finding to verify: …""`. These were already answered
  from the dataroom. Your job is to independently look up the same fact from the
  public source and report **what the public source says**, so it can be compared
  against the document's claim.

## How to work

- For a fill item, first consider its `specific_public_source`. If the uploaded
  documents already contain the value, use that and cite the document.
- For a cross-check item, report the public-source finding on its own merits -
  do **not** just echo the prior document finding. If the public source agrees,
  say so and cite it. If it points to something materially different (e.g. the
  documents describe a company as a dedicated BESS specialist but public sources
  show a broad renewables developer), report the public-source picture plainly -
  the difference will be surfaced to the reviewer.
- Record the state (found / probably_found / not_found), the source you used, a
  short quote where the evidence is text, and your confidence.
- Never invent a value. If neither the documents nor your knowledge of the named
  source gives a reliable answer, mark it not_found.

## Live web search - ONLY when the tool is actually available

Some requests in this phase include the **web_search** tool; most do not.

**If the web_search tool is available in this request**: use it. Search for the
confirmed selling party, site and project at the named `specific_public_source`
(or a good public source for that fact), read what you find, then report the
public-source value. Prefer a cited public source over your own recollection.
When you have what you need, finish by calling `record_findings`. Traceability
is mandatory for every web-searched finding:

- **evidence_source**: list the actual URLs/domains you used (e.g.
  `northdata.com/...; 50hertz.com/...`) - never a dataroom document name. The
  reviewer must be able to open your sources and check them.
- **reasoning**: say what each source stated and how you concluded - including,
  for a cross-check, whether the public record confirms, contradicts, or is
  silent on the prior document finding. "Prior finding verified" alone is not
  acceptable; name what verified it.
- If your searches came up empty, record not_found and list in `reasoning`
  which sources you tried.

**If the web_search tool is NOT available in this request**: answer from the
uploaded documents and your background knowledge only. Never claim you checked,
searched or verified anything against a public source - you didn't. Never cite a
URL you have not actually read in this request. For a cross-check item you can
still note whether the prior finding is consistent with your background
knowledge, but say exactly that ("consistent with background knowledge"), not
that it was verified.

> Note for developers: live web search is enabled via `web_search: on` in this
> file's frontmatter (set it to `off` as a cost kill-switch). Only cross-check
> (`c`) items are eligible, further narrowed during the trial by an explicit
> allowlist (`_WEB_SEARCH_TRIAL_ITEMS` in engine.py). Each item runs in its own
> call with **no documents attached** and the search tool in **direct mode**, small
> max_tokens, one resume, a hard wall-clock deadline that cancels the stream and
> records an explicit not_found. Turning the flag off reverts the whole phase to
> knowledge-only with no other change.

```

```
## Items to fill in this phase

Fill each checklist item listed below - one finding per item, and only for these
items. Each line is `item_id | factor / topic | extract: <what to extract> | look
at: <public source>`, optionally followed by a CROSS-CHECK instruction.

{{items}}

## Output

Call the `record_findings` tool with one entry per item you were given.
```

D.3. API Payloads, Call Trace, and Cost

This section specifies what leaves the machine per call type, what is recorded of every call, and what a complete run costs. Cost matters to this research for one reason: the scalability condition of Chapter 2 and the time-saving requirement (US-02) both fail if screening support is expensive per prospect. The measurements below show that it is not, by a wide margin.

What is sent to the API, per call type

Table D.5 specifies every call type the prototype makes. Each request consists of document blocks, one instruction block, a forced tool schema, and generation settings, assembled as described in the previous section.

Table D.5: API payload specification per call type.

Call	Model	Max out. tokens	Documents attached	Instructions	Tool (forced)	Web search
Site detection	Haiku 4.5	1 500	Teaser files	general + site_detection	record_sites	no
Quick extraction	Haiku 4.5	25 000	Teaser files (top-level)	general + quick + 52 item lines + feedback	record_findings	no
Data-room triage	Haiku 4.5	(small)	None (file listing as text: paths, sizes, pages)	triage instruction	classification schema	no
Full extraction	Sonnet 4.6	50 000	Data room, or routed selection $\leq$ 90 pages / 28 MB per request, $\leq$ 220 pages per run	general + full + in-scope item lines + feedback	record_findings	no
Gaps, knowledge pass	Sonnet 4.6	32 000	Routed selection, $\leq$ 120 pages per run	general + gaps + item lines (cross-check suffix carries prior value) + anchors + feedback	record_findings	no
Gaps, live web item	Sonnet 4.6	(small)	**None**	general + gaps + one item line + anchors	record_findings	yes

The per-call trace

Every real API call is recorded in the prospect's trace file (`sent_prompts.json`): the exact assembled instruction text, the list of attached documents, the tool schema, the model, the token cap, the timestamp, and the measured token usage including cache reads and the stop reason. A run log accumulates the same data per historical run, which is the basis of the per-prospect cost accounting reported below and in Chapter 8. The trace means that for any finding in any evaluated prospect, the complete request that produced it can be reopened and inspected. The reproducibility claim of this thesis rests on that property rather than on the reproduction of model outputs, which remain version-dependent.

Cost per prospect

Input cost scales with the pages sent, and the routing budgets of Section 7.3 are the ceiling mechanism. Output cost is capped per call by the skill file's token limit. Table D.6 gives the order of magnitude per prospect in the measured configuration. The demonstration campaign of Chapter 8 measured $10.58 in total API cost for nine prospects, with deep runs between $1.68 and $4.51 and quick runs between $0.06 and $0.20 (Table 8.2). Set against the specialist hours a screening record supports, the marginal cost of a run is three to four orders of magnitude smaller.

The cost picture carries two qualifications.

Table D.6: Order-of-magnitude API cost per prospect.

Step	Typical cost
Quick extraction (teaser)	\$0.06–0.20
Data-room triage (once per file set, cached)	about \$0.02
Full extraction (routed, at most 220 pages)	\$1.5–2.0
Gaps: allow-listed live web checks	about \$0.5
Gaps: knowledge pass (at most 120 pages)	\$0.5–1.0

Prompt caching, within a run and beyond. The API is stateless, but statelessness does not preclude caching: a request can mark a prefix for reuse, and identical prefixes in subsequent requests are then read at roughly a tenth of the input price [7]. The prototype marks the attached document block in this way, so the sequential calls of a single run (batching retries, per-factor routing) do not re-pay the full document cost; the cache reads are visible in the per-call token usage of the request trace. The larger lever appears only at organisational scale. The fixed opening of every extraction request (the shared rules and the 152 item lines) is identical across prospects, so a shared deployment would pay for it rarely instead of once per prospect. Each request stays independent and nothing from one prospect reaches another; the saving comes from reusing identical instruction text, not from remembering content.

A snapshot, not a forecast. These prices are a snapshot of one provider's list prices in mid-2026, in a market whose providers are widely reported to operate at a loss. They indicate the order of magnitude of usage, not a forecast of what organisation-wide operation will cost over the lifetime of an adoption. The conclusion drawn from them in Chapter 8, that price is not the binding constraint on adoption, survives even a substantial price increase, because the distance to analyst-hour costs is measured in orders of magnitude rather than percentages.

D.4. Illustrated Prototype Demonstration

This section illustrates the prototype from the perspective of the human reviewer, following the complete path of a single prospect through the tool. To avoid disclosing confidential material, a fictional prospect named *Project Meridian* was created purely for this demonstration: a BESS project in the Netherlands, accompanied by a single teaser document. The screenshots are presented in the chronological order in which a reviewer would encounter them. The pipeline behind these screens, its three phases and its review gates, is specified in Chapter 7; the account below stays with what the reviewer sees and does.

Creating a prospect

A prospect is created on the *New Prospect* page (Figure D.1) by giving it a name, selecting the country, and uploading the relevant files. The tool distinguishes documents by their position in the prospect folder: files placed at the top level are treated as the teaser, while a seller's data room is recognised by its nested sub-folders. In this example only a teaser PDF is supplied, so the walk-through covers the teaser-based part of the pipeline.

Running the quick extraction

Once created, the prospect appears in the *Prospect Overview*, which lists every prospect together with its headline figures and its current position in the pipeline. Starting the quick extraction runs it in the background, and the status column reports progress (Figure D.2). The quick phase uses the cheaper model, reads only the teaser, and fills the small set of headline items that a teaser is expected to contain; it also detects whether the teaser describes one site or several.

When the phase completes, the overview updates to invite review and surfaces the extracted headline values (here 200 MW and 800 MWh) directly in the table (Figure D.3).

Reviewing the quick data

The review screen (Figure D.4) opens with a location button that links to the prospect on a map at the most precise level available, a set of headline cards, and a summary of how many checklist items were

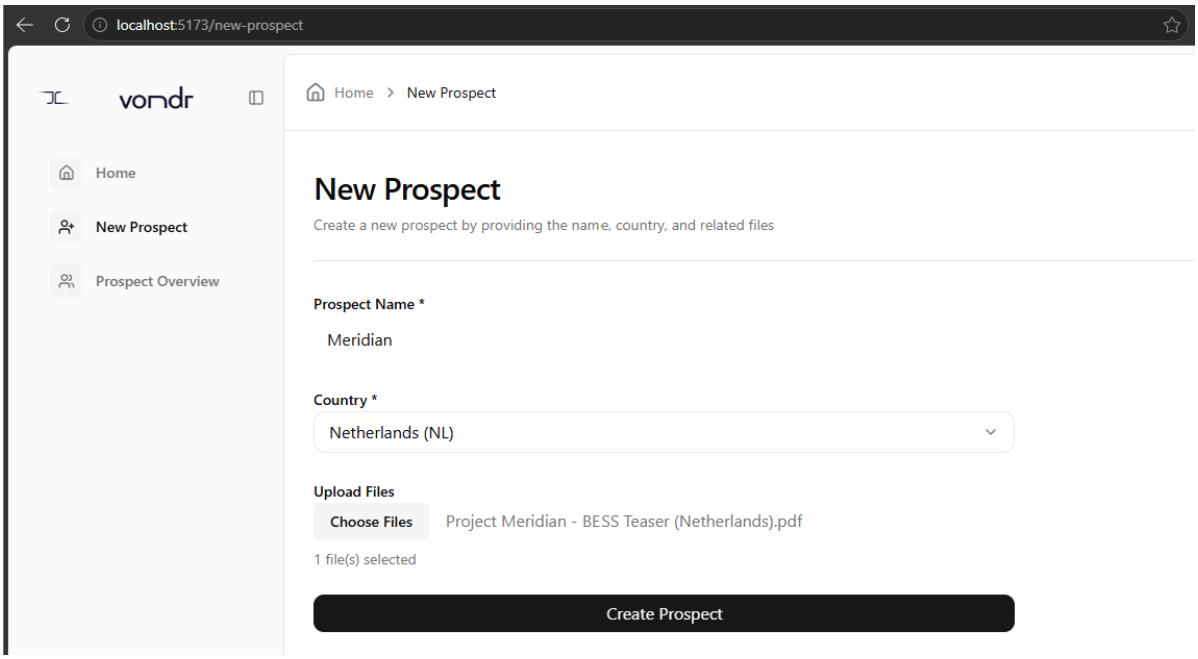

Figure D.1: Creating the prospect: name, country and the uploaded teaser document.

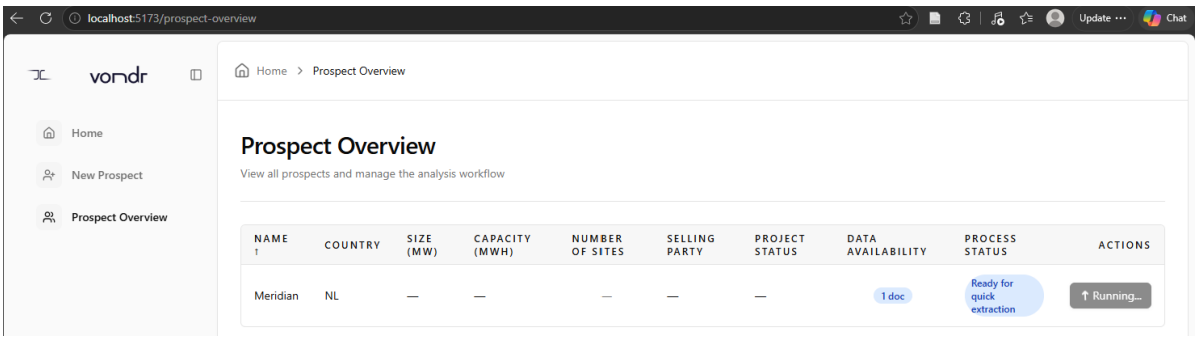

Figure D.2: The prospect overview while the quick extraction runs in the background.

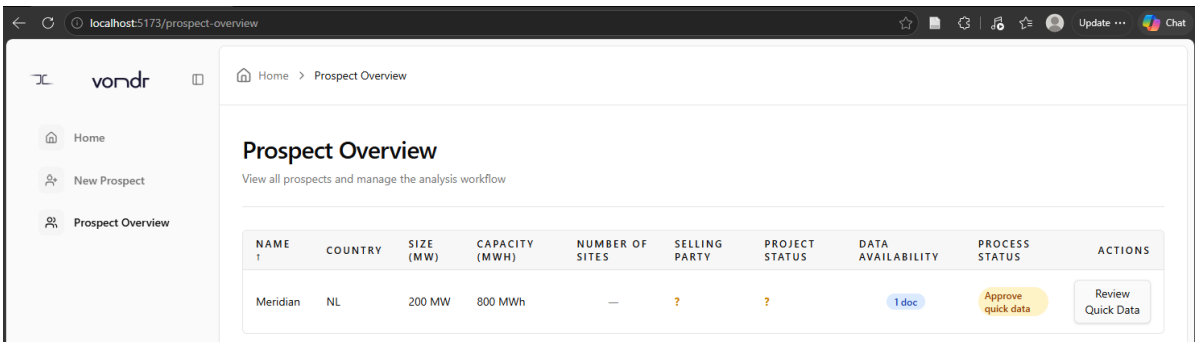

Figure D.3: After the quick extraction, the headline values are populated and the prospect is ready for review.

found, judged probable, not found, or not searched for in this phase. A *not searched for* item is one that falls outside the current phase's scope, which is kept distinct from an item that was searched for but genuinely absent.

Critical anchor values must be explicitly confirmed before the phase can be approved (Figure D.5). For each such value the reviewer may accept the machine's reading, change it (either to correct a misread or to supply a value the model could not find), or mark it as unknown at this stage. These values are treated as critical because later phases build upon them.

The full set of items is shown in a table grouped by factor (Figure D.6). Each row records the finding, the reasoning behind it, the evidence state, a confidence level, and the source together with a short quotation. Quotations are automatically checked against the cited document, and a badge indicates whether an exact textual match was found. The plot coordinates in this example illustrate a derived value: the model converted the teaser's degrees-minutes-seconds coordinates into decimal form and recorded the conversion in the reasoning field.

When every critical value has been confirmed and every probable finding has been given a verdict, the phase can be approved; alternatively it can be declined with feedback (Figure D.7).

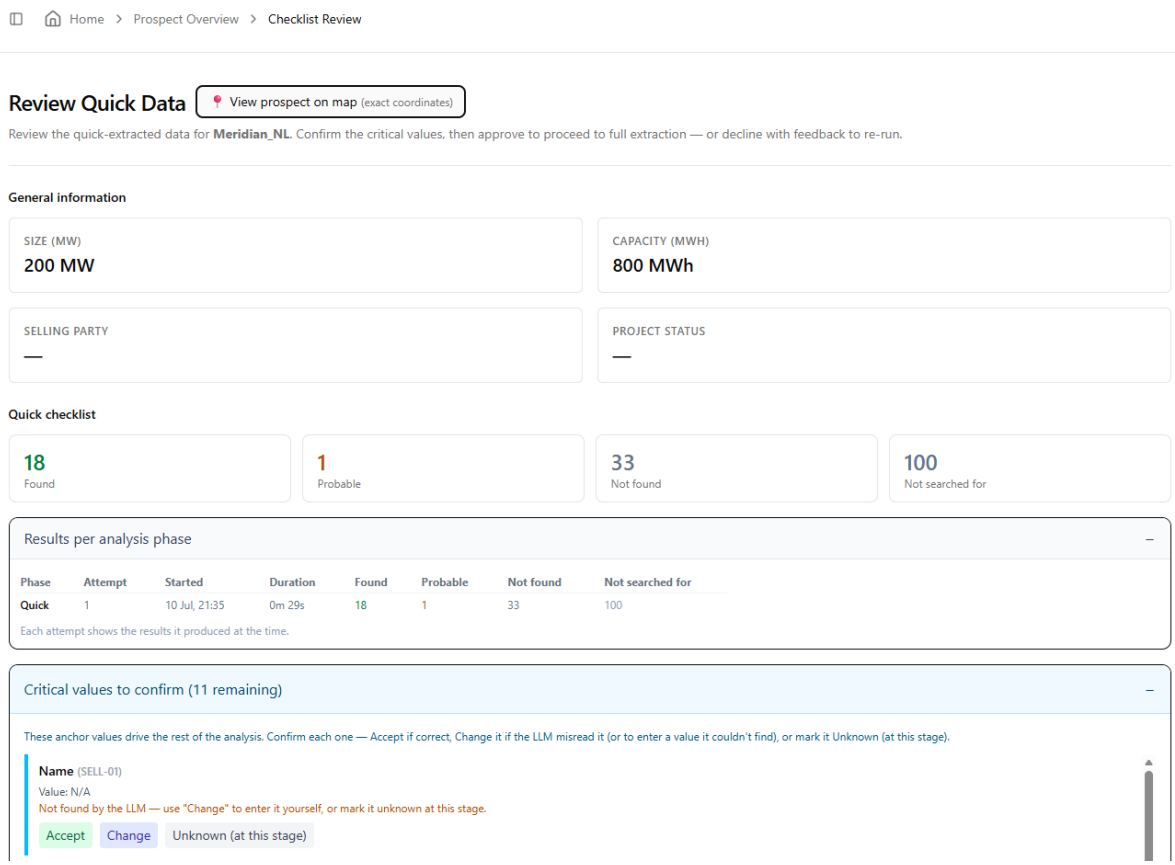

Figure D.4: Top of the quick-data review: the map link, headline cards, the per-state summary, and the critical-values gate.

Declining with feedback and re-running

If the reviewer judges the extraction inadequate, the phase can be declined with a written explanation (Figure D.8). In this example the reviewer notes that the current land use was missed and adds a rumour that the site might be expandable. This feedback is not merely stored: it is inserted into the prompt of the subsequent re-run, so the model is told explicitly what the reviewer expected.

The re-run opens with the previous feedback displayed at the top of the screen, and the *results per analysis phase* panel now records both the first and the second attempt, each with its own timing (Figure D.9). The effect of the feedback is visible in the table: the current land use is now reported as

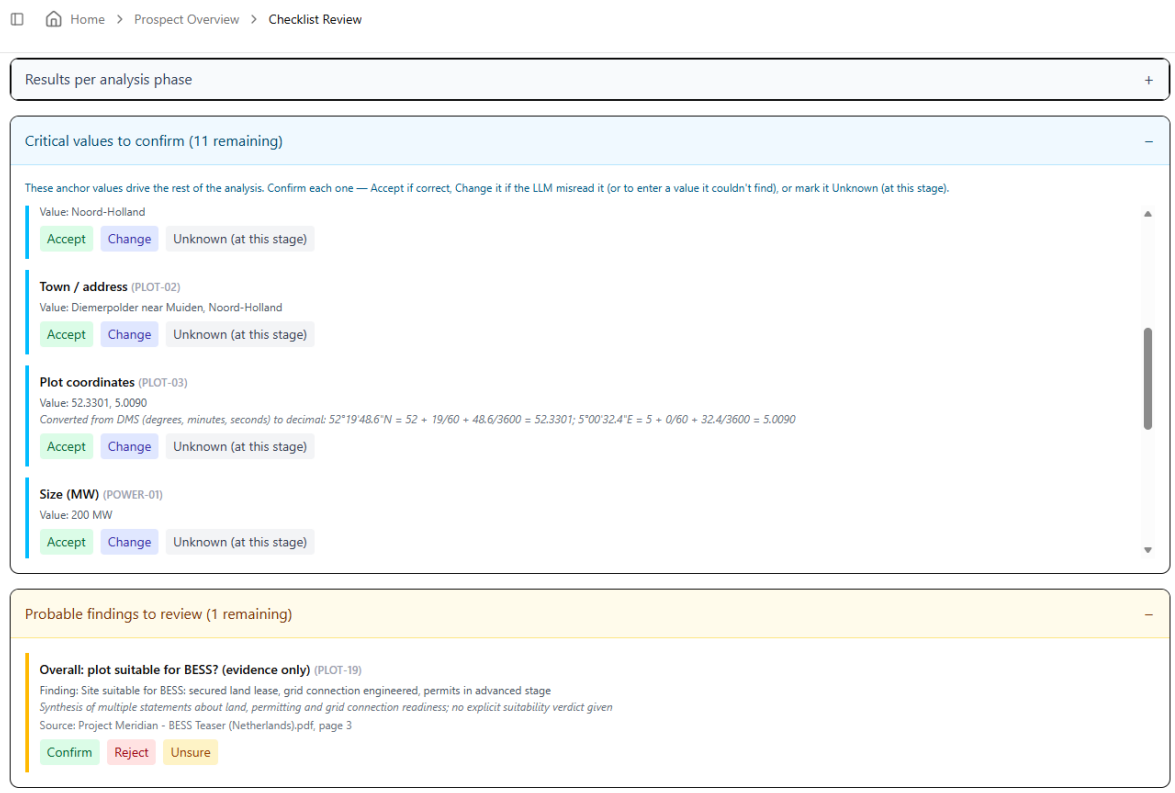

Figure D.5: Confirming critical anchor values: each may be accepted, changed, or marked unknown at this stage.

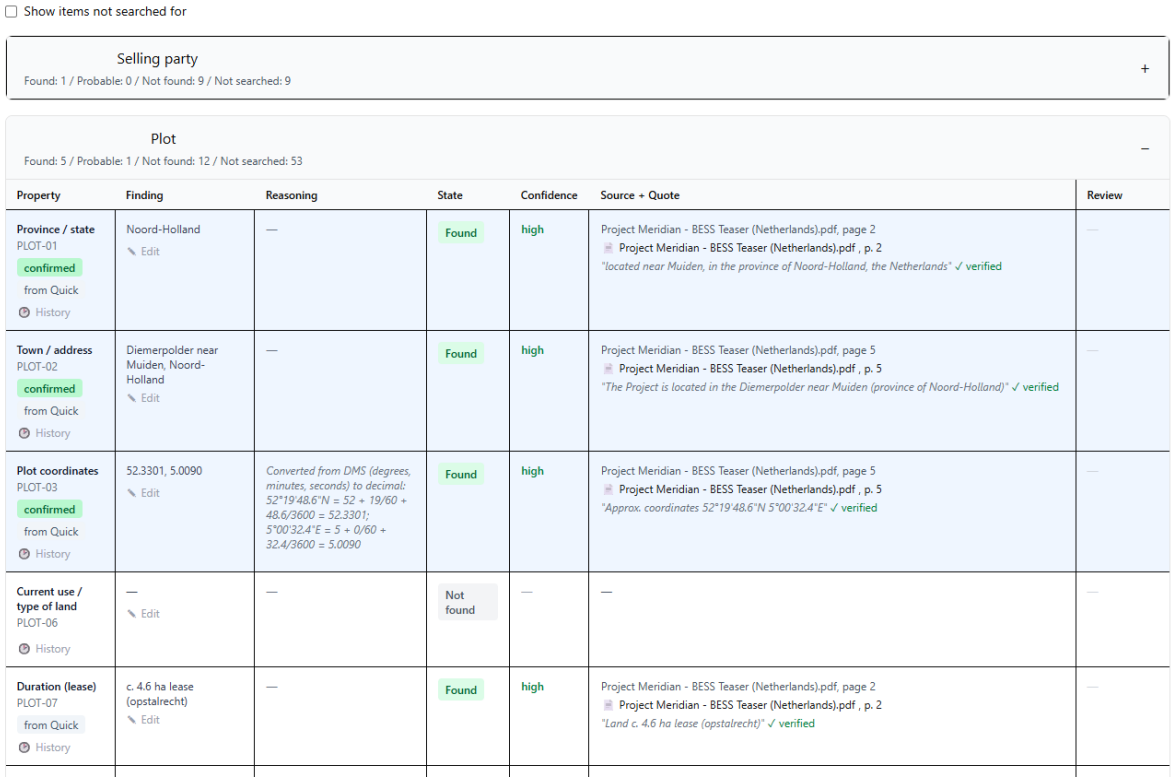

Figure D.6: The checklist table grouped by factor, showing findings, reasoning, evidence state, confidence, and verified source quotations.

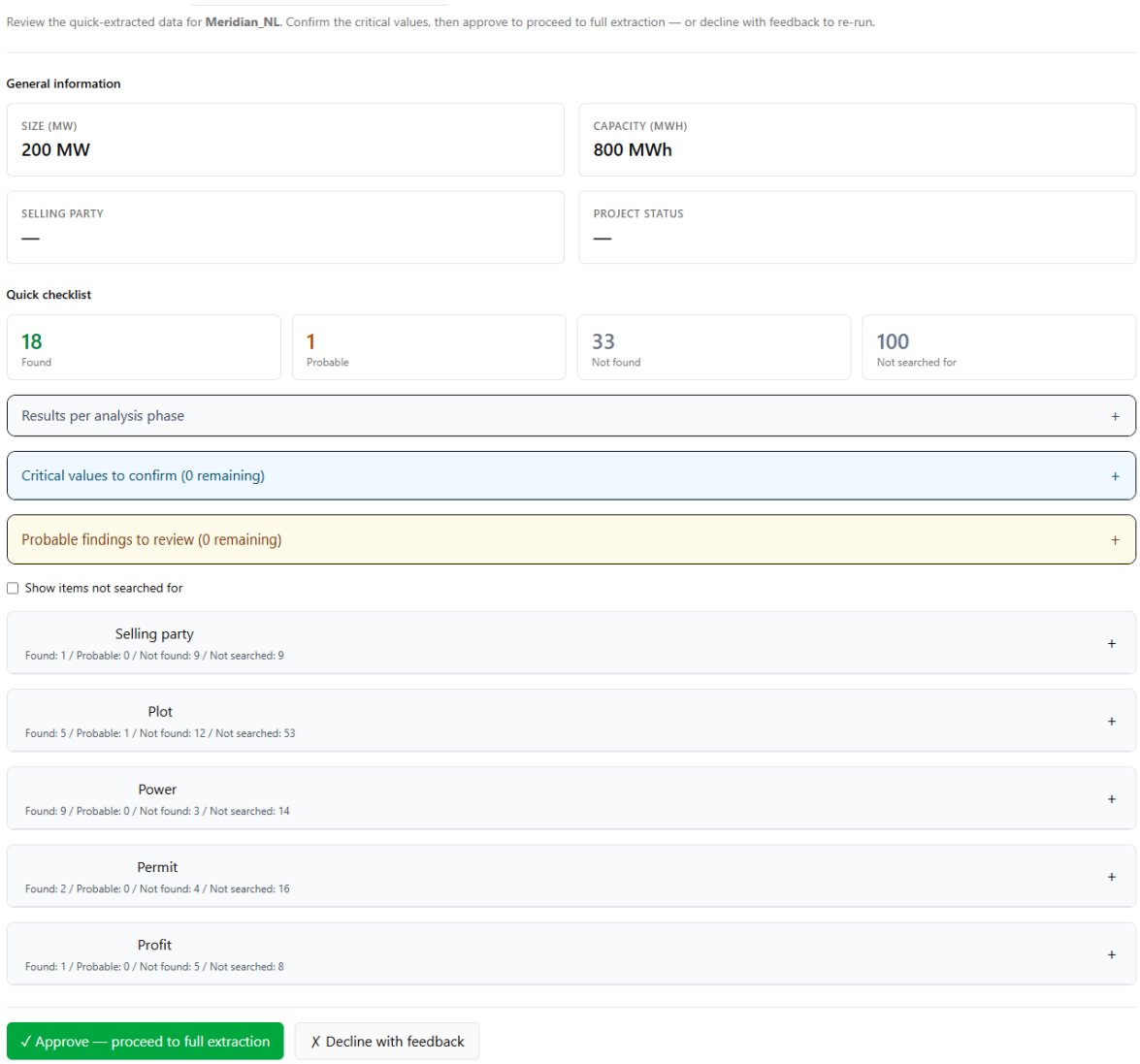

Figure D.7: Once the gates are cleared, the reviewer may approve the phase or decline it with feedback.

agricultural, whereas the first attempt had left it unanswered (Figure D.10).

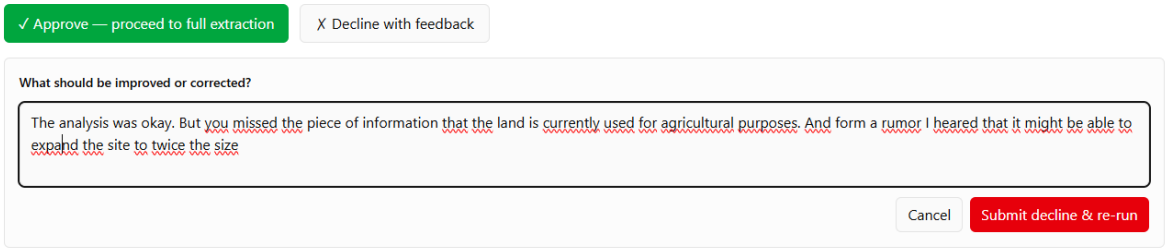

Figure D.8: Declining the phase with written feedback, which is passed to the model on the re-run.

Reviewing the full extraction

After the quick data is approved, the full extraction phase runs against the complete document set and answers the detailed checklist. Its review screen follows the same pattern but adds the cross-phase comparison (Figure D.11). The per-phase results panel now lists the quick attempts and the full attempt together, and three panels organise the reviewer's work: differences to resolve, critical values to confirm, and probable findings to review. A difference is raised whenever a later phase produces a value that disagrees with what was already on record; quantitative disagreements beyond a set tolerance, and any disagreement on a critical value, must be resolved before approval.

The checklist table gains a dedicated review column on the right for recording a verdict on each finding (Figure D.12). The source column continues to show the automatic quotation check, distinguishing a verified quotation from one for which no exact textual match could be found. Selecting a cited document opens it in a viewer within the application, scrolled to the exact page of the citation, so the reviewer can inspect the evidence without leaving the tool (Figure D.13).

Completing the pipeline

Once the full extraction is approved, the prospect moves into the third and final phase: *fill the gaps*. This phase serves two purposes. First, it provides a last opportunity to fill any checklist item still marked as not found. Second, and more distinctively, it cross-checks a small, deliberately limited selection of items (among them the identity and structure of the selling party, the identity of any sell-side adviser, and the status of the grid connection) against publicly available sources, using a server-side web-search capability of the language model. This selection is kept narrow by design, since live web search carries a materially higher cost and latency per item than reading an already-uploaded document. For every item that is checked in this way, the outcome is recorded explicitly as one of four states: confirmed by the public source, differing from the document, filled where the documents were silent, or no public data found. This distinction matters to the reviewer, because it separates a genuine independent verification from a value that merely happens to agree with what the documents already said, and it separates a considered absence of public information from a check that was never attempted.

The review screen of this phase follows the layout already shown for the quick and full extractions, with two additions. Items that were cross-checked against publicly available sources carry a badge showing the outcome of that check. Any contradiction between a document finding and a public source is raised as a difference, in the same manner as a contradiction between two extraction phases, and the reviewer resolves it exactly as illustrated above by choosing which value is carried forward. The screens of this phase are therefore not reproduced here.

When the fill-the-gaps phase has been approved, the checklist itself is considered complete: every item has been attempted according to its scope, every critical value has been confirmed, and every disagreement between phases, and between the documents and the public record, has been resolved by the reviewer. The prospect then leaves the checklist-filling pipeline illustrated here and becomes available for the next stage of the process, in which the completed checklist is used to assess the prospect as a whole, as described in Section 7.1.

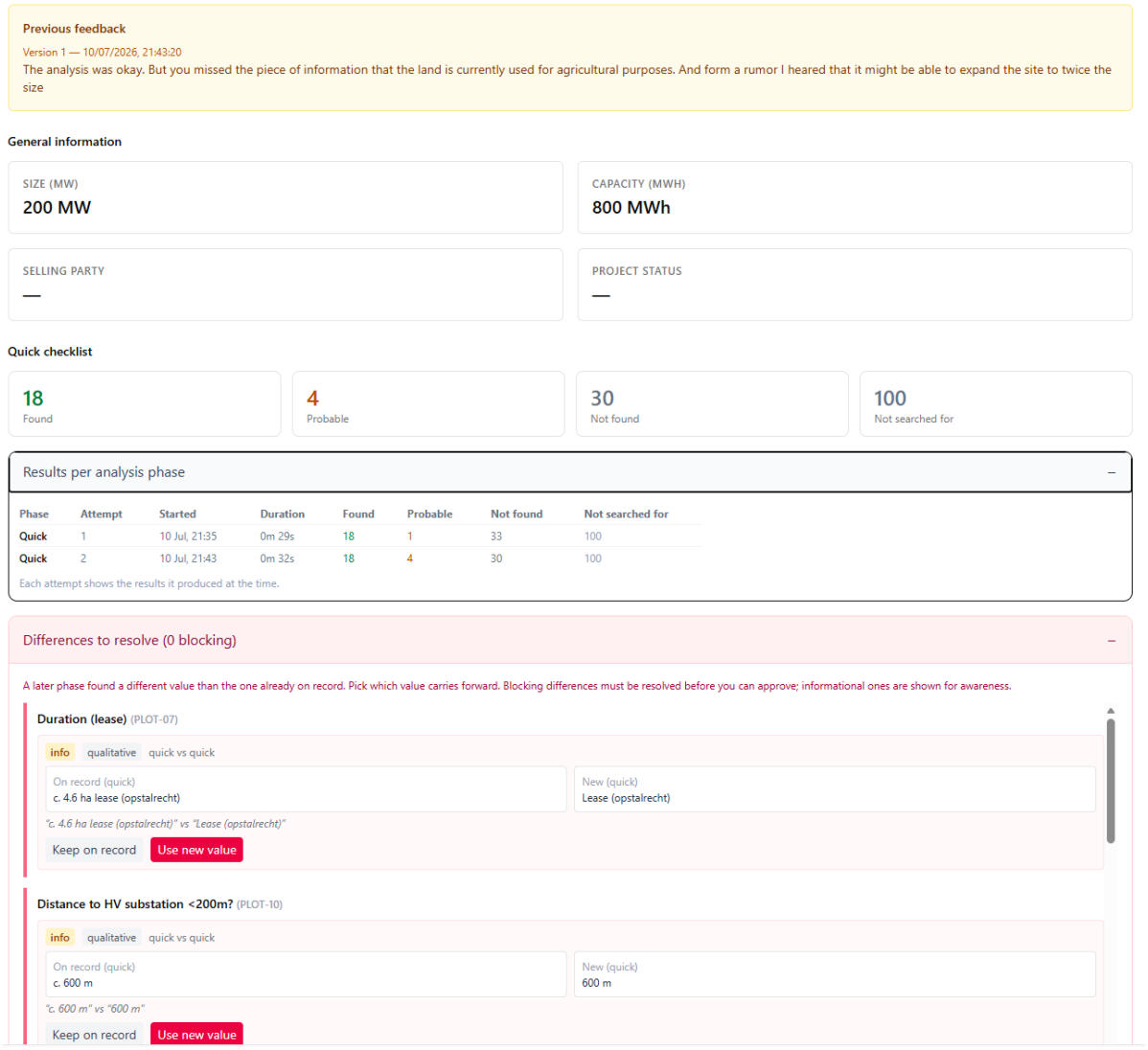

Figure D.9: The re-run shows the previous feedback and records the new attempt alongside the first in the per-phase results.

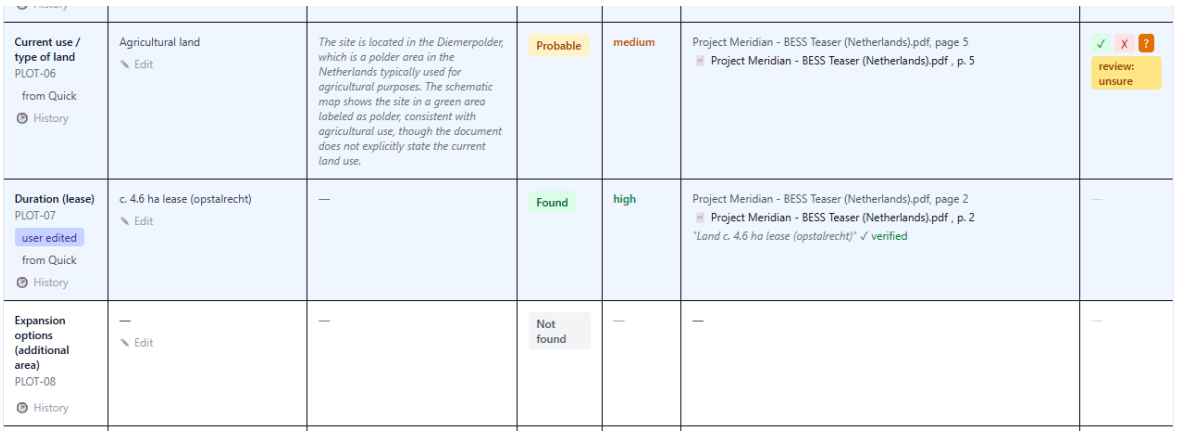

Figure D.10: The feedback has been acted upon: the current land use is now identified, whereas it was previously absent.

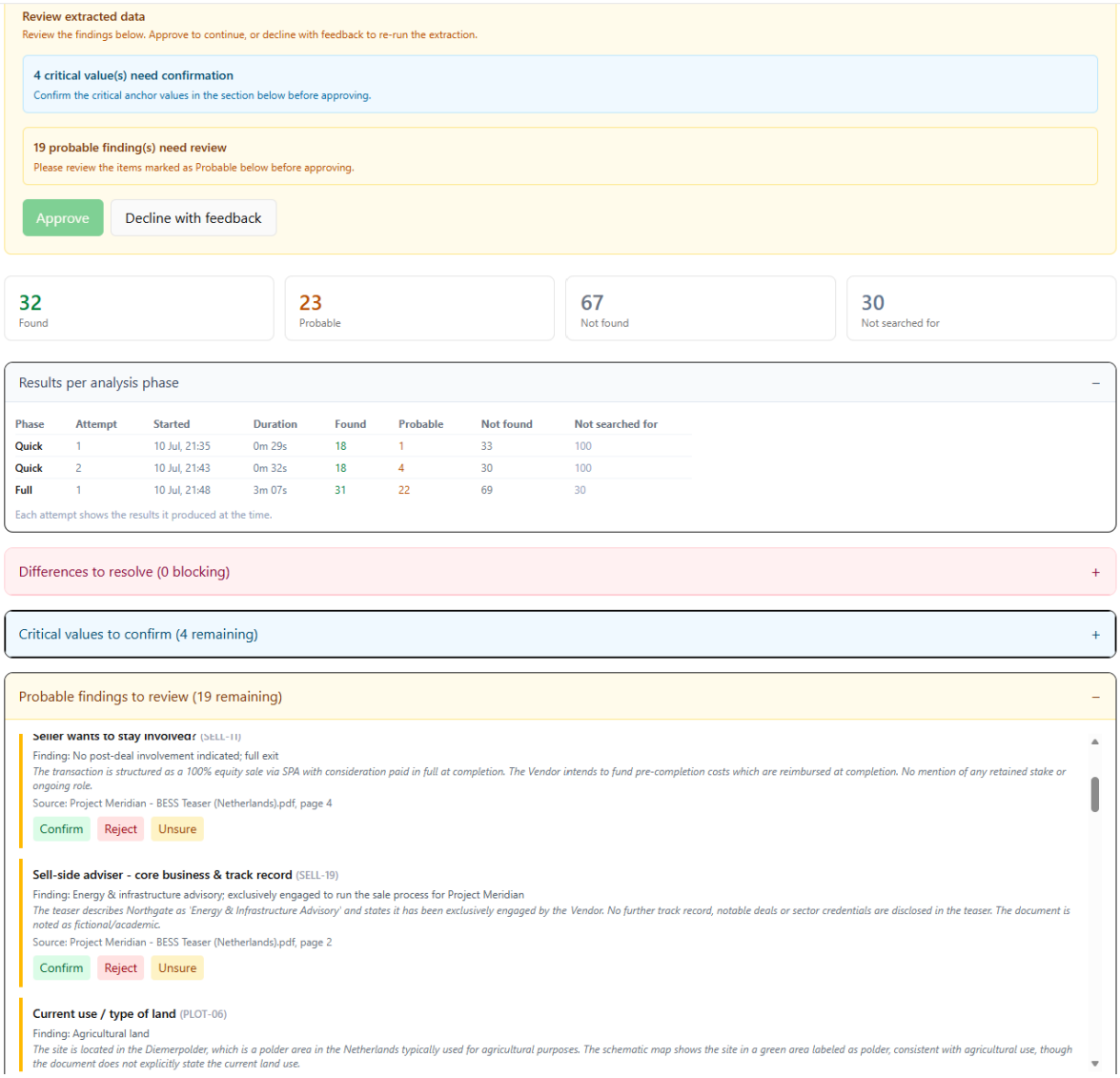

Figure D.11: The full-extraction review: the per-state summary, the per-phase results across attempts, and the panels for differences, critical values, and probable findings.

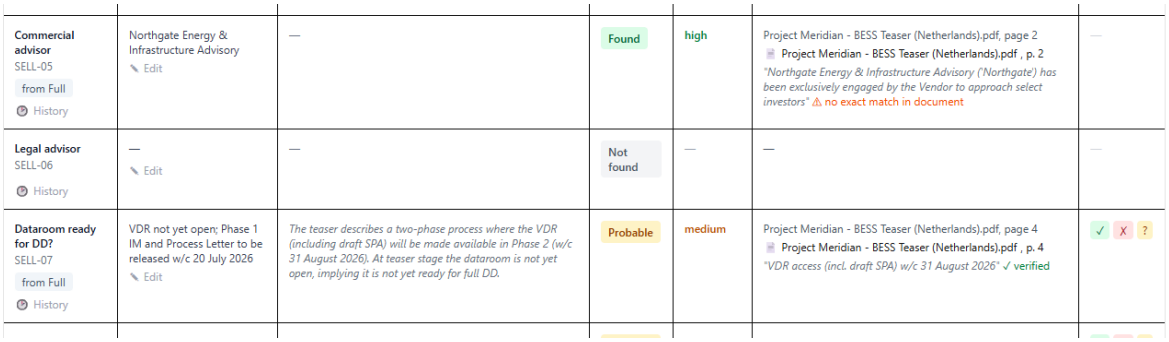

Figure D.12: The full checklist table, with the review column on the right and the automatic quotation-verification badges in the source column.

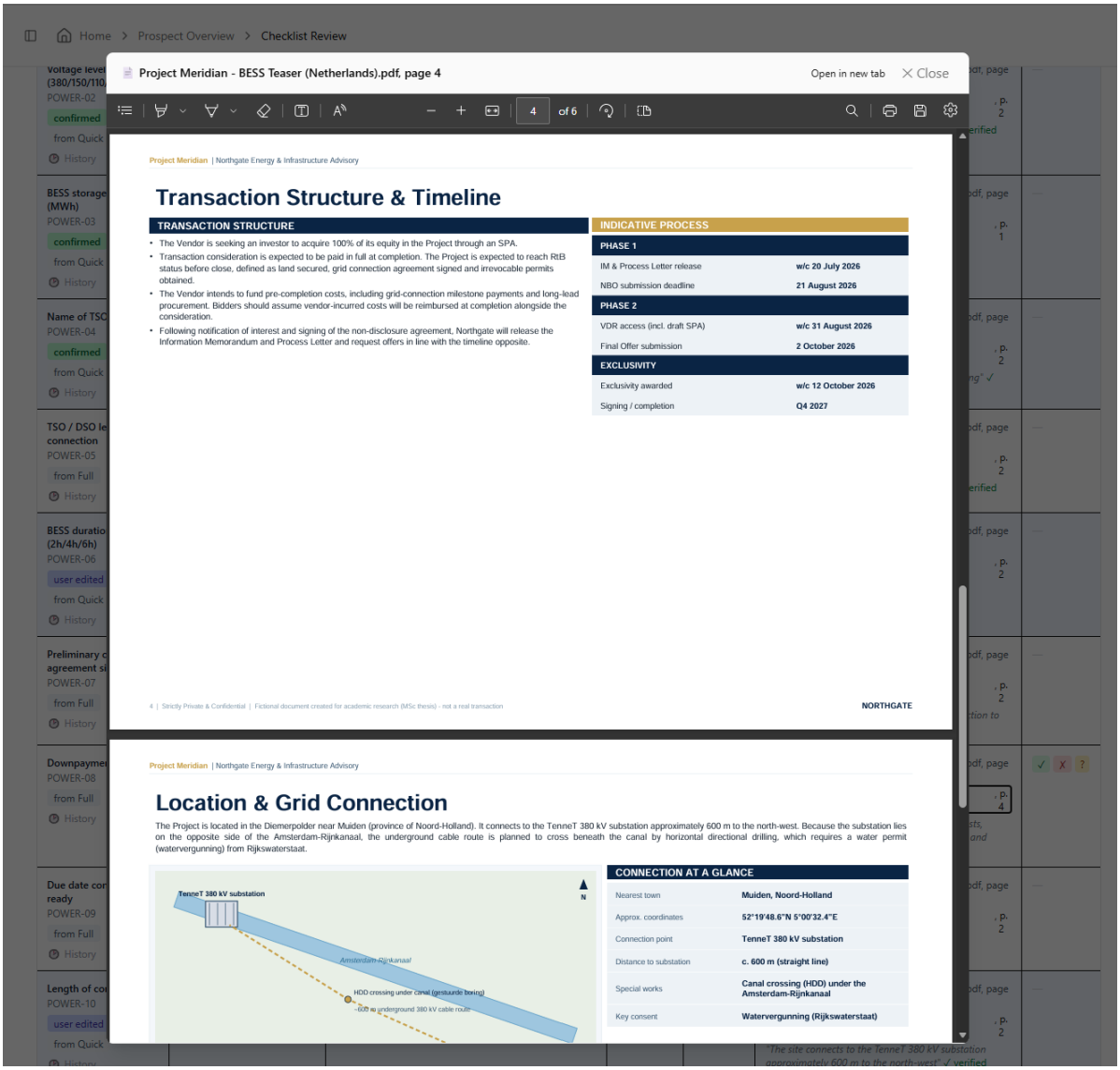

Figure D.13: Selecting a citation opens the source document inside the application at the cited page.

D.5. Development Log

This section condenses the development history of the prototype, as evidence of the iterative skills-definition loop of Section 3.5.1 and of the design cycle discussed in Chapter 9. Development proceeded in twenty logged phases between the project start and the evaluation freeze. Phases mark shifts in focus rather than formal milestones. Phases 0 and 1 predate systematic logging and are reconstructed, while everything from phase 2 onward is backed by the session history, the prompt traces, and version control. Table D.7 gives one line per phase; the closing subsection states the cross-cutting principles that the iterations converged on, which Chapter 9 distils into design principles.

Table D.7: The twenty development phases of the prototype.

Phase	Focus	Principal changes and lessons
0	Origins	Start from a company-provided FastAPI + Vue application template; first prospect pages built by hand; early extraction of a small hardcoded parameter list.
1	Checklist schema and prompt externalisation	The 152-item version-6 schema with per-phase scope flags and critical markers; all prompts moved out of code into editable instruction files; documents sent as native files; finding separated from reasoning. Lasting principle: only assembled prompt text ever leaves the machine.
2	Cost and correctness of extraction	A per-factor call split that re-paid the document cost five times was replaced by adaptive single-call batching with truncation-triggered halving; per-phase model selection and prompt caching introduced; parser hardened on real edge cases.
3	Review unification	One shared review component at every phase gate; phase-scope logic centralised; distinction between not found and not searched made faithful to true phase scope.
4	Multi-site portfolios	Site detection spawns child sub-prospects sharing the parent's documents; identity and path logic centralised.
5	Full prompt traceability	The last inline prompt moved to a file; the per-prospect trace introduced, recording the exact model, settings, documents, instructions, and tool schema of every call.
6	Dead-code cleanup	Legacy services, views, and unused checklist fields removed after tracing; a clean base for the central redesign.
7	Persistence fix	Review edits were silently lost in one mode; reconceived so that every commit persists immediately in all phases.
8	Cross-phase comparison (central redesign)	Immutable per-phase snapshots plus a merged master record; per-item history; conflict detection with 1%/5% tolerances and full both-sides records; sticky user ownership; blocking conflicts gate approval; the gaps phase becomes a cross-check that corroborates but never overwrites documents.
9	Live web search cost incidents	The most instructive arc: successive token leaks traced through a fan-out flood, an unstoppable server-side loop, and a watchdog that abandoned a still-billing stream, to the root cause: the search tool's default execution mode runs a billed server-side sandbox. Fixed by direct mode, one item per call with no documents attached, a one-search cap, and a hard deadline with real stream cancellation. Lesson: bound spend structurally.

Phase	Focus	Principal changes and lessons
10	Per-item history	The phase-by-phase timeline, current-value highlight, and per-item conflict log exposed in the review interface.
11	Honest internet checking	Per-item verification badges with consulted URLs; the label “cross-check” reserved for genuine live searches; the prompt split so tool-less calls may never claim public-source verification or cite unread sources.
12	Seller versus adviser	Real prospects conflated the selling party with the sell-side adviser; item definitions sharpened and dedicated adviser items added, both web-verified; the live trial widened to seven items.
13	Reviewer feedback loop	Decline feedback, stored from the start, is now injected into every re-run prompt, with per-item rejection reasons appended; a live test caught one phase bypassing the wiring, settled by the trace.
14	Prompt file audit	Editor notes separated from model-facing text; language, number, and date normalisation rules added; contradiction rule added; a citation audit revealed the model had never seen filenames, fixed by titling every document block.
15	Formats and citations	Spreadsheet, CSV, and e-mail conversion at upload; structured citations (exact document and page) added to the findings schema, powering a click-to-open document view at the cited location.
16	Data-room scale	The triage-and-route mechanism of Section 7.3: local inventory, one cached classification call over the file listing, and budgeted high-relevance-first routing per factor, with every dropped file logged and not-consulted notes on affected absences.
17	Review ergonomics	A batch of gate-usability improvements to the shared review component.
18	Design-flaws review and hardening	A systematic pass over known weaknesses; extraction made fail-loud (a phase whose calls all fail raises instead of writing a silent all-not-found result); routing budget allocation made fair across factors.
19	Guardrails	Data-room presence guard before expensive runs; per-run instrumentation freezing each attempt’s counts; stale critical confirmations reopen on new evidence; non-destructive rejection with reasons sent to the model on re-run; precision-aware map links.
20	Freeze	Checklist validated; dead code swept; the architecture, prompt-design, and data-model documentation written; a results aggregator producing the campaign tables of Chapter 8; the development set separated from the evaluation runs, with one data room held out.

Cross-cutting principles

Six principles recur across the phases and were treated as invariants once established: every model-visible byte is inspectable, and editable content lives in exactly one place; human decisions are never destroyed by machine output; claims are verified rather than assumed, with regressions diagnosed from the trace; spend is bounded structurally, not aspirationally; honesty outranks impressiveness, so the system never claims verification that did not happen and never drops work silently; and statelessness is explicit, so everything the model should know is deliberately injected and therefore appears in the trace. Chapter 9 states these as the design principles this thesis contributes.

D.6. From Prototype to Production: Security and Data Handling

The prototype documented above was built as a research instrument, and two classes of requirement separate it from a deployable system. Both were identified during the security review of 6 July 2026 (Section 9.5) and neither was implemented or tested within this research, which bounds the claims of Chapter 9 accordingly.

Adversarial inputs and a red-team test set

Documents in a seller's data room are supplied by an interested counterparty, so they must be treated as untrusted input rather than as neutral evidence. The prototype constrains the model on the output side, because the forced answer schema leaves no channel for free-form instructions to be echoed back, but it applies no input-side defence. Before confidential material is processed at organisational scale, the following test set should be run and its results recorded:

1. **Document prompt injection.** PDFs carrying instructions to the model in body text, in white-on-white or zero-size text, in image captions, and in document metadata.
2. **Hidden instruction persistence.** Injected text that attempts to alter later phases through the anchor values, the feedback field, or the merged record.
3. **Exfiltration attempts.** Instructions directing the model to place data room content into a finding, a source field, or a search query.
4. **Malformed and hostile files.** Corrupted PDFs, deeply nested archives, spreadsheets with formula payloads, and files whose declared type differs from their content.
5. **Cross-prospect leakage.** Runs executed in sequence to confirm that no content, cache entry, or confirmed value of one prospect can reach the request of another.
6. **Tool abuse.** Attempts to induce unbudgeted web searches or to defeat the page, call, and deadline limits.

A finding from any of these tests is a blocking defect rather than a refinement, because the failure mode is the disclosure or corruption of counterparty-confidential information.

Data handling requirements beyond file-based storage

The prototype stores each prospect as plain JSON in its own folder (Section 7.4). That choice serves inspectability and traceability during research, and it satisfies none of the handling requirements that confidential transaction data carries in production:

- encryption at rest and in transit, with key management separated from application access;
- role-based access control at prospect level, so that access follows the deal team rather than the installation;
- append-only audit logging of reads, edits, confirmations, and exports, retained independently of the record it describes;
- concurrency and versioning, so that simultaneous reviewers cannot silently overwrite one another;
- backup and recovery with a stated recovery point, and integrity verification of restored records;
- retention and deletion per counterparty, so that material can be removed when an NDA requires it, including from backups;
- observability of the traffic between the harness and the model provider, as the security review specified.

These requirements are stated here because they belong to the transfer of this design into use, and because the absence of any of them would make the record less trustworthy than the manual practice it replaces.

E

The LLM-Integrated Analysis Process

This appendix documents the complete process design for LLM-integrated prospect analysis, of which the thesis implements the foundational block in depth (the scope decision is motivated in Section 7.1). It serves two purposes. For the reader of this thesis, it preserves the parts of the conceptual design that frame the implemented prototype but are not themselves implemented, so that the main text can concentrate on what was built and evaluated. For the BESS developer and for researchers who continue this work, it is the development roadmap: it consolidates the known limitations of the current prototype (Section E), and it describes how the structuring layer grows into a fully LLM-supported investment risk assessment (Section E). The research-level limitations of the thesis remain in Section 9.7; what is collected here are the artefact-level refinements that a BESS developer would address before organisation-wide use.

From structuring to assessment: the full analysis loop

Figure E.1 shows the LLM-supported assessment loop. In the loop, defined skills instruct the model to analyse a prospect against the risk factors of Chapter 5; the model's output is reviewed against expert knowledge of the case; and discrepancies between output and expert judgement are fed back into the skills, so that successive rounds converge towards output that domain experts regard as substantively correct. The loop is a feature of the development process, not of a deployed instrument.

This diagram was the methodological starting point of the design phase of this research. The design work itself then showed that the loop presupposes something the context did not yet provide: a reliable, structured record of the prospect for the model to assess. The thesis therefore descended one level and implemented the structuring layer first (Chapter 7). The loop remains valid at that level, where it was applied as the skills-definition loop of Section 3.5.1, and it returns in full at the assessment level once its inputs exist. Read in that light, the figure is no longer a description of this thesis's method; it is the process recommendation for the next stage of development.

The place of this loop in the wider process is defined by the four-block design of Section 7.1 (Figure 7.1): the risk-factor catalogue, the filled prospect checklist, the market memory, and the organisational assessment memory converge on the assessment step, and feedback on assessed prospects flows back into the organisational memory. The blocks that remain to be built are described below.

Process designs of the memory-building blocks

Two recurring back-end blocks supply the assessment with the knowledge that neither the data room nor the checklist contains. Figure E.2 shows their process designs.

The *market overview* block condenses public market information and commissioned market reports into a compact market memory file, refreshed on a recurring basis (bi-weekly to monthly) rather than per prospect. Skills retrieve and analyse the public sources; a framework skill condenses the results into the memory file; and a human review at configuration time, rather than per prospect, checks the con-

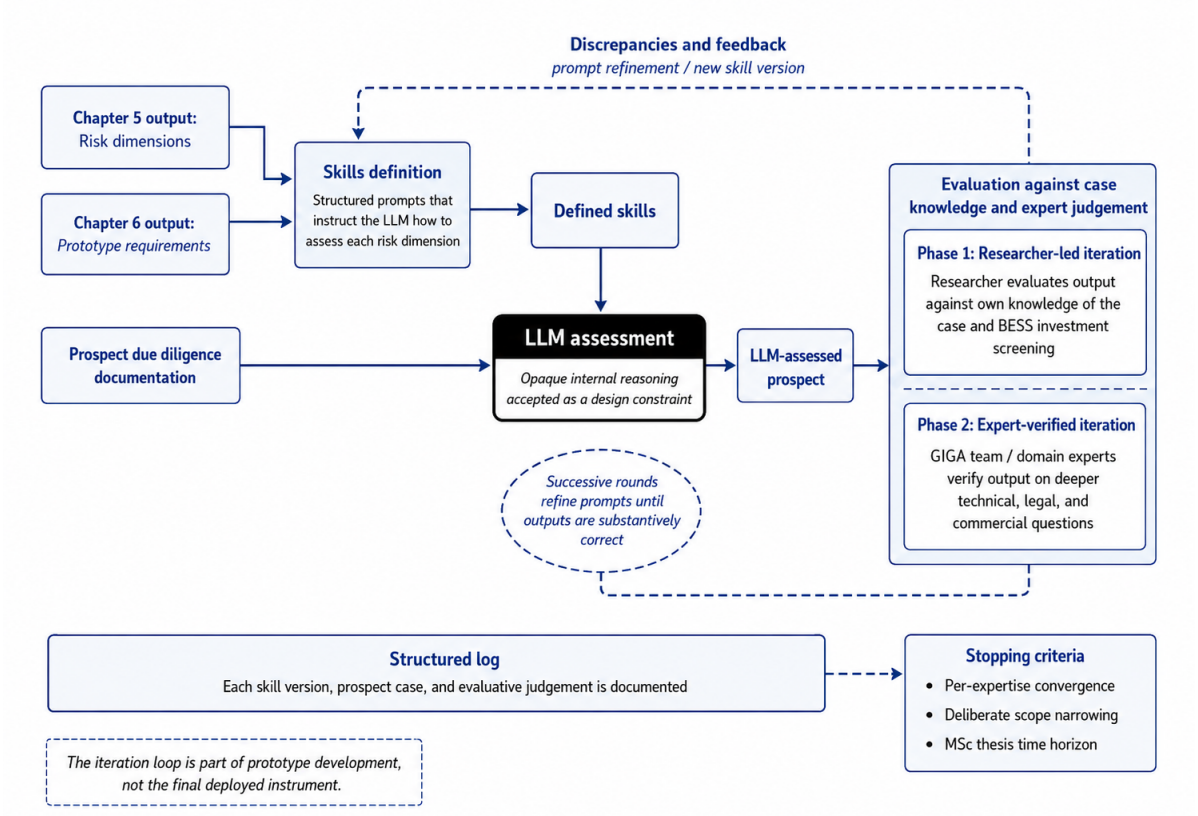

Figure E.1: An LLM-supported assessment loop process recommendation

densation before the assessment consumes it. The evaluation gave this block its first evidence of being load-bearing rather than optional: several permit items are named after jurisdiction-specific statutory instruments, and the walk-through of Appendix C.3 showed that the functional question (for example, whether an environmental impact assessment is required) must be separated from the jurisdiction-specific instrument that answers it. Resolving that mapping per country is precisely the role of the market memory.

The *assessment definition* block captures how the organisation itself evaluates projects. Prior cases, recurring decision questions, meeting transcripts, and feedback on assessed prospects are condensed into an organisational assessment memory, so that the assessment step reflects the organisation's own decision logic and improves with use. This block is also where the institutional-memory requirements of Chapter 6 (FR-05, FR-18) find their long-term home: logged records of past prospects and rejection reasons become working material for the organisation's assessment definition, not only an archive. One design caution applies from the start: a memory that preserves expertise preserves outdated assumptions and bias with the same fidelity, so the block needs explicit mechanisms for challenging, versioning, and retiring captured rules, not only for capturing them.

Known limitations of the current prototype

The development log (Appendix D.5) and the evaluation of Chapter 8 together surfaced a set of artefact-level limitations. They are the expected condition of a prototype at this stage of development rather than findings about the approach, which is why they are consolidated here instead of in the main text. Read as a list, they are the first work package for a BESS developer before organisation-wide implementation. Table E.1 states each limitation, its cause, and the direction of the fix.

Table E.1: Artefact-level limitations of the frozen prototype, their causes, and fix directions.

Limitation	Cause	Direction of fix
Explicit non-values (“TBD”) oscillate between *found* and *not found*	Evidence model has no placeholder rule	One-sentence rule in the shared instructions
Confirmed critical values do not constrain the full phase; derived figures go stale	Anchor injection wired to the gaps phase only	Inject anchors into every later phase; recompute derived values on anchor change
Semantically identical values raised as conflicts; no way to combine two partial answers	Textual comparison of qualitative findings; single-choice conflict resolution	Semantic-equivalence filter; merge action in the review
Legally weighty formats skipped (binary Q&A spreadsheets, word-processing files, saved web pages, shapefiles)	Format converters not built	Convert-to-text pipeline for Office binaries and web archives
Format-skipped files trigger no per-item not-consulted note	Honesty notes wired to budget drops only	Extend the not-consulted mechanism to format skips
Triage assigned every candidate document the same relevance	Relevance rubric does not discriminate	Sharpen the rubric; use folder taxonomy and document-type priors
Routing budget bounds document coverage (44 of 232 in the audited run)	Fixed page budget as cost ceiling	Budget sensitivity testing; spend budget on instruments over summaries via document typing
Seller-composed summaries preferred as cited sources	No document provenance classes in the design	Type documents by author; rank or exclude summaries
Sub-sites of portfolio teasers inherit portfolio-level findings	No site-specific prompting	Per-site item lines for multi-site prospects
The same fact phrased differently across phases reads as a different answer	No normalisation of qualitative findings	Canonical phrasing rules; semantic comparison
Quote verification fails on scanned or heavily tabular text	Text-layer extraction limits	Structured PDF processing separating text from visuals
Gaps-phase logs do not preserve which in-scope items were not attempted	Aggregator records returned items only	Log the full scope decomposition per run
Borderline evidence states flip between identical runs (3.8% of items)	Residual run-to-run variance of the model	Tighter item specification; repeated runs with majority vote for critical items

Development roadmap

The evaluation converts the remaining work from a wish list into an ordered roadmap. The ordering follows the dependency logic of the four-block design and the priorities the practitioners themselves expressed. Items marked (P) originate from the practitioner sessions; items marked (E) from the instrumented evaluation and walk-throughs.

1. **Re-specify the checklist instrument.** In this design each item is compiled directly into the extraction prompt, so item quality is prompt quality. The work comprises: tightening the extraction definition of every ambiguous item, starting from the instances of Table C.5 (E); rules for explicit non-values and for not-applicable answers, complementing the three evidence states (E, P); posing jurisdiction-functional questions and resolving the country-specific instrument via the market memory (E); criticality tiers, so that a defined subset of items is always human-verified regardless of model confidence (P); per-item trusted-source whitelists for public verification (P); item prioritisation by decision influence (P); and an explicit channel for observations that fit no item, so that standardisation does not narrow attention to the risks already in the schema.
2. **Harden the extraction pipeline.** Work through Table E.1 from top to bottom; the anchor wiring, the document typing, and the format coverage carry the most due-diligence weight. Iterative reruns as new documents arrive should become a first-class operation (P), together with a side process that drafts questions to the selling party from the gaps the extraction identifies (P).
3. **Broaden independent verification.** Live public-source checking currently covers an allow-listed subset of items; extending it requires the per-item source specifications of step 1. The discriminating experiment identified in the walk-through comes first: withhold the seller's overview documents and re-run, to establish whether source preference is governed by the prompt or by the routing (E). A process variant proposed in the group session belongs here as well: an internet-augmented, teaser-only pathway, in which public sources fill gaps directly after the quick phase, before any data room exists (P). This variant serves the earliest and most common screening situation and is a practical interest of the BESS developer; it was scoped out of the present research.
4. **Build the memories.** Implement the market-overview and assessment-definition blocks of Section E, in that order; the market memory unlocks the jurisdiction mapping of step 1, and the assessment memory requires the logged prospect history that steps 1–3 accumulate.
5. **Design the assessment stage, gated by the known-baseline study.** With structured records, a market memory, and an organisational memory in place, the assessment step of Figure E.1 can be designed: risk factors weighed against the structured facts, with the human review shifting from doing the analysis to controlling it. Before any organisation relies on it, the known-baseline study defined in Section 9.8 (running the pipeline on completed prospects and measuring accuracy and time saved) is the decisive gate.
6. **Harden for production.** The security findings of the BESS developer's review stand at the end of the roadmap because they concern deployment rather than design: prompt-injection shielding, input sanitisation, structured PDF processing that separates text from visuals before merging, zero-trust sandboxed document handling, observability of the traffic around the harness, and the governance decision between public API access and models run on the organisation's own infrastructure (Section 9.5).

The academic counterparts of this roadmap (the known-baseline study, document-set ablation, item specification as an experimental variable, and the evaluation of a fourth evidence state) are developed as future research in Section 9.8.

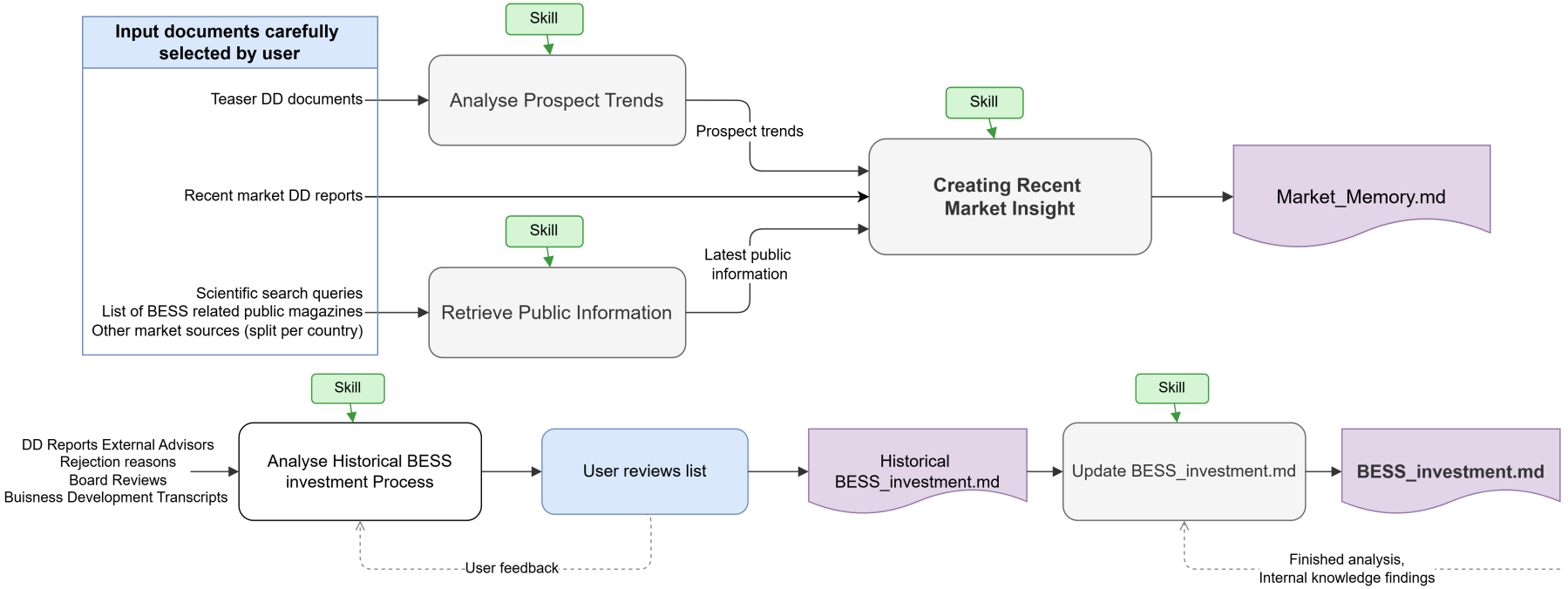

Figure E.2: Process designs of the two recurring memory-building blocks. Top: creating the market overview. Bottom: defining the BESS assessment at the BESS developer.

F

Full Risk-factor catalogue

This appendix presents the full version of the risk-factor catalogue introduced in Chapter 5. The columns extend Table 5.1 with two further fields: the industry or sector of the supporting source(s), and the assessment method as given in the supporting source where one was provided. “Interviews” in the industry column denotes that the factor is also supported by the expert interviews described in Section 3.4.3. The table is sorted by identifier and presented in landscape to accommodate the column widths.

Identifier prefixes encode the academic category as in Table 5.1. The *Industry/sector of source(s)* column indicates the original empirical context of each supporting source, alerting the reader to factors whose transferability to the BESS context relies on an extrapolation. The *Assessment method* column records the assessment approach where a supporting source provided one. “Interviews” in the industry column denotes that the factor is also supported by the expert interviews described in Section 3.4.3.

Table F.1: The full risk-factor catalogue of 52 investment risk factors.

ID	4P	Risk factor	Industry/sector	Sources	Assessment method
CON-02	Power	Supply chain disruption (equipment delivery)	Construction, BESS	[93, 8, 94]	Manufacturing slot reservation early in project cycle; force majeure clauses; dual-sourcing where feasible [8].
CON-06	Plot	Heavy-transport and logistics accessibility	Interviews	P2, P3, P4, P7	Bring forward to early screening: check road network and bridges via public infrastructure data.
CON-07	Profit	Project schedule slippage and consequential cost overruns	Interviews, Hydropower, power plants, construction, infrastructure	[85, 51, 16, 32]; P1, P2, P4, P5, P6, P8, P9	Add time/planning sub-dimension to every risk indicator (P6: cost × quality × timing).

continued on next page

Table F.1 continued from previous page

ID	4P	Risk factor	Industry/sector	Sources	Assessment method
CPT-01	Profit	Offtaker credit quality (TSO, utility, corporate PPA buyer)	BESS, renewable energy	[11, 94, 61]	Credit rating review; parent-company guarantees; counterparty concentration limit; consensus-based DD for fintech/green [61].
CPT-04	Profit	Prospect developer information quality	M&A, BESS, Interviews	[32, 11]; P1, P2, P5, P9	Outside-view reference class benchmarking [32]; independent technical and legal DD; site visit.
CPT-05	Profit	Counterparty corporate hygiene (KvK, KBO, registries, blacklists)	Interviews	P1, P2, P5, P9	Highly automatable: integrate KvK/KBO/Handelsregister + EU-/OECD blacklist + news scan.
CPT-06	Plot	Information completeness ('what is missing')	Interviews	P1, P2, P4, P5, P6, P9	Tool must distinguish 'absent' from 'present'; flag absence as separate category, not as low score.
ECO-01	Profit	High initial capital expenditure (CAPEX)	Renewable energy, ESS, Interviews	[1, 81, 19]; P1, P4, P6, P7, P8	NREL ATB cost projections as benchmark [19]; reference class forecasting against completed projects [32].
ECO-03	Profit	Electricity revenue curve uncertainty (price level and spread compression)	ESS, renewable energy, Interviews, BESS	[63, 6, 41]; P2, P3, P4, P5, P6, P8	Stochastic price paths (GBM, jump-diffusion) in financial model [63]; price scenarios from national TSOs/PBL.
ECO-06	Profit	Macro-economic exposure (interest rates, lithium and battery prices)	Hydropower, renewable energy, Interviews, BESS	[85, 6, 94, 19]; P2, P6, P8	Sensitivity of NPV to WACC over 100-300 bps; fixed-rate hedge analysis.
ECO-07	Profit	Inaccurate economic assumptions (CAPEX, OPEX, inflation)	Hydropower, construction, Interviews	[85]; P8, P9	Escalation clauses in EPC; inflation-linked debt service coverage analysis.
ECO-11	Profit	Rising grid connection fees and tariffs	BESS (NL specific), Interviews	[99, 10]; P1, P2, P3, P5, P6, P7, P8	Project-level modelling of kW-contract and kW-max under current and proposed ACM tariff structures; comparison with DE grid-fee exemption baseline [99].
ECO-13	Profit	Bankability threshold compliance	Interviews	P2, P4, P6, P8	Reverse-engineer screening criteria from financial-close requirements.
ECO-14	Profit	Financial-model assumption sensitivity (discount rate, acquisition-vs-financing mismatch)	Interviews, Renewable energy	[6]; P6, P8, P9	Standardised assumption list with type tag (generic/regional/project-specific) and source attribution.
ENV-01	Plot / Permit	Noise impact	Hydropower, BESS, Interviews	[85, 91]; P3, P4, P7	Acoustic modelling against local noise ordinance; noise barriers / setback design.
ENV-02	Plot	Flood / hydrology risk	Hydropower, BESS	[85]	Hydrological site assessment; avoidance or defined setback (BESS checklist practice).
ENV-03	Plot / Permit	Ecological / biodiversity constraints	Hydropower, infrastructure, BESS	[85, 55, 91]	Ecological survey and EIA; BBN propagation to permit denial and delay [55].
ENV-04	Plot	Climate / weather extremes	BESS, hydropower, Interviews	[85, 94, 20]; P2, P4, P7	HVAC sizing validated for local climate extremes; OEM warranty conditions on temperature range [94, 103].
GRD-01	Power	Grid connection unavailability / waiting list	BESS, renewable energy	[99, 47, 29, 10]	TSO/DSO congestion maps (NL: Netbeheer Nederland); ATR85 or time-bound transmission contracts [10]; prioritisation framework eligibility review.
GRD-02	Power	Non-firm connection curtailment risk	BESS (NL, UK, FR), Interviews	[99, 10]; P1, P2, P3, P4, P5, P6, P7, P8, P9	Co-optimisation model over DAM/FCR/aFRR with curtailment hours as constraint [99]; sensitivity to curtailment frequency.
GRD-03	Power	Grid connection technical constraints (voltage, capacity)	Hydropower, BESS, Interviews	[85, 37, 89]; P1, P3, P7, P9	TSO pre-application feasibility study; connection cost estimation against similar precedent projects [37].

continued on next page

Table F.1 continued from previous page

ID	4P	Risk factor	Industry/sector	Sources	Assessment method
GRD-04	Power	Uncertainty in grid operator processing time	BESS, renewable energy, Interviews	[99, 47]; P1, P2, P3, P4, P6, P8, P9	Option-value framing of connection queue position; scenario analysis over delay tails.
GRD-06	Power	Grid connection contract status (legal certainty)	Interviews	P1, P2, P3, P4, P5, P6, P8, P9	Implement as ordinal maturity ladder per jurisdiction with binary 'signed-or-not' overlay; primary fatal-flaw indicator.
LEG-01	Plot	Land title and rights defects	Hydropower, renewable energy, BESS	[1, 11, 94]	Legal DD including Kadaster title verification; lease term vs asset life analysis.
LEG-02	Permit	Incomplete permitting or consents	Renewable energy, BESS	[11, 91, 74]	Permitting red-flag DD checklist ([91], SEAI manual provides BESS-specific list); consent schedule review.
LEG-03	Profit	PPA / tolling / capacity agreement terms	BESS	[94, 8]	Mark-up review against standard heads of terms ([94], WB PPP guidelines Appendix A); Morgan Lewis / DWT benchmark negotiation points [8].
LEG-05	Profit	Information asymmetry vs prospect developer (M&A)	M&A broadly, BESS, Interviews	[32, 11]; P1, P2, P4, P5, P6, P9	Outside-view benchmarking against comparable projects [32]; structured red-flag DD across financial, legal, technical, strategic streams [11].
LEG-06	Profit	Tax structure and transaction risk	M&A, BESS, Interviews	[11]; P2, P8, P9	Tax DD opinion by specialist; transaction structuring review.
MKT-01	Profit	Data availability and reliability for market/consumer behaviour	Renewable energy, BESS	[1, 94]	Proxy markets / cross-jurisdictional price benchmarks; publicly available TSO portals (ENTSO-E, Regelleistung, TenneT).
MKT-02	Profit	Market design change (ancillary service market evolution)	BESS	[86, 41, 94]	Track market operator consultation; stakeholder engagement with BRPs [37].
MKT-03	Profit	Value-stacking constraints (regulatory)	BESS (EU)	[86, 31]	Member-state-specific stacking review ([86], EU analysis); validate stacking assumptions against Grid Code and TSO rules.
MKT-04	Profit	Counterparty concentration / offtake dependency	BESS, M&A broadly, Interviews	[11, 94]; P8	Counterparty credit rating review; PPA/tolling contractual protections; diversification via multi-service provision.
MKT-05	Profit	Country-level BESS market viability	Interviews	P1, P2, P3, P6	Country-level pre-screen before any project-level assessment.
PLT-01	Profit	Adverse legislative, tax or change-in-law risk	Renewable energy, hydropower, ESS, BESS	[1, 85, 6, 25, 90, 31, 94]	Stochastic scenario modelling with policy shock scenarios ([75], multi-stage stochastic); real options to value deferral [63].
PLT-02	Profit	Policy retraction (withdrawal of support)	ESS, Interviews	[90]; P2, P3, P5, P6, P7, P8	Asymmetric scenario modelling that weights retraction heavier than new subsidy [90].
PLT-06	Profit	Regulatory change as 'unknown uncertainty' (information gap)	Interviews	P2, P4, P6, P8	Most automatable AI use case: scrape public regulatory docs continuously per active market.
POL-01	Permit	Permitting and licensing delay or refusal	Renewable energy, hydropower, BESS	[1, 91, 94]	Storage Readiness Level scoring of permitting status ([4], analogue); SEAI Ireland consenting manual provides process map and typical timelines [91].
POL-02	Power	Uncertainty in grid extension plans and TSO/DSO planning	Renewable energy, BESS	[1, 29, 47]	Review TSO 10-year development plans; ACER monitoring reports [29]; PBL/TenneT forward scenarios for NL [47].
POL-03	Permit	Inadequate or unclear asset classification for storage	BESS	[31, 86, 94]	Regulatory tracking of national transposition; review of Network Code and tariff decisions.
POL-04	Permit / Plot	Restrictive land-use / zoning rules for BESS	BESS, renewable energy, Interviews	[1, 91]; P1, P2, P3, P4, P5, P6, P9	Early zoning review; engagement with municipality; site screening against local spatial plans.
POL-05	Permit	Permit issuing authority identification	Interviews	P1, P2, P3, P5, P9	Automated check: identify jurisdictional authority from public policy documents at start of screening.

continued on next page

Table F.1 continued from previous page

ID	4P	Risk factor	Industry/sector	Sources	Assessment method
POL-06	Permit	Municipal cooperation signal (principebesluit/Vorbescheid)	Interviews	P1, P2, P3, P4, P5, P9	Document presence of principebesluit/Vorbescheid as a quantifiable maturity indicator.
POL-07	Permit	Country and regional permit procedure variation	Interviews	P1, P2, P3, P4, P5, P9	Country/region module needed; cannot be a single generic ruleset.
POL-08	Permit	Municipal inexperience with BESS as new asset class	Interviews	P2	Contextual benchmark: cross-municipality energy-transition activity as proxy for permit-process speed.
RES-01	Profit	Revenue resource uncertainty (market price spread volatility)	BESS, ESS	[63, 41, 6]	Monte Carlo simulation over historical DAM spreads ([63], real options approach); scenario analysis over renewable penetration trajectories; stress testing to 2030 capacity build-out.
RES-02	Profit	Ancillary service price and availability uncertainty	BESS, Interviews	[41, 99, 81, 86]; P1, P2, P7, P8	Sensitivity analysis with progressive saturation assumptions; review of TSO ancillary procurement rules and volume caps; EU benefit-stacking regulatory tracking [86].
SOC-01	Permit	Local community opposition / social acceptance	Hydropower, renewable energy, BESS, Interviews	[1, 85, 55, 43]; P1, P3, P7	Early community engagement; social impact assessment; BBN for socio-environmental risk propagation [55].
SOC-03	Permit	Public safety perception (post-fire-incident)	BESS	[43, 73]	Community risk assessment per NFPA 855 Annex guidance; transparent communication of HMA findings to AHJ and community [73].
TEC-04	Power	Thermal management and fire / thermal runaway	BESS	[37, 77, 73, 97, 43]	UL 9540A large-scale fire testing at cell/module/unit/installation level [97]; NFPA 855 Hazard Mitigation Analysis (HMA) including evaluation of initiation, propagation, gas composition, deflagration hazards; Thermal Runaway Propagation Prevention (TRPP) systems [73].
TEC-06	Plot	Site geotechnical / soil conditions	Hydropower, construction, BESS, Interviews	[85, 51, 93]; P3, P4, P7	Geotechnical investigation as part of feasibility; BBN causal modelling [51] showed inadequate soil investigation as critical driver of cost overrun in power plants.
TEC-09	Plot / Power	Cable route feasibility (length, parcel count, route width)	Interviews	P1, P3, P7, P9	Geospatial calculation: shortest route + Kadaster parcel count = quantitative risk proxy.
TEC-10	Plot	Land area suitability for planned capacity	Interviews	P4, P7	Automatable as a minimum-area check given plot polygon and target MW.

G

Full Requirements Specification

This appendix provides the full specification of the 41 requirements summarised in Chapter 6. Each requirement is listed with its IEEE 29148:2018 sub-type, full description, MoSCoW priority, robustness score across the eligible interview subset, and the interviewees who raised it. The interview codes (P1–P9) correspond to the roster in Appendix B.

For brevity in the main thesis text, the tables in Chapter 6 present only abridged descriptions and omit sub-type, robustness, and source attribution. The present appendix is the definitive reference for the requirements specification.

Functional/Performance Requirements

Table G.1: Functional/Performance requirements.

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
FR-01	Risk scoring	Per-dimension risk classification with traffic-light output	Produce a green/amber/red (or 4-level) risk classification per 4P dimension and per cross-cutting dimension, derived from probability × impact assessment per indicator.	Should	Strongly corroborated	P4, P6, P8

continued on next page

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
FR-02	Risk scoring	Composite project quality scorecard with three-axis scoring per element (cost, quality, timing)	For each 4P dimension, score elements on three independent axes: cost/price, quality (substantive completeness), and timing/planning. Aggregate to dimensional score and overall project quality number.	Could	Role-specific	P6, P8
FR-03	Risk scoring	Phase-relative threshold calibration	Score the same indicator differently depending on project maturity phase (early/mid/ready-to-build). A missing element acceptable in early phase may be a fatal flaw at financial close.	Won't	Role-specific	P4, P6
FR-04	Risk scoring	Fatal-flaw flagging independent of aggregate score	Some conditions are project-killers regardless of other strengths (e.g. no signed grid offer, no land rights, hostile permitting authority). Flag these separately and override aggregate score.	Should	Strongly corroborated	P1, P2, P3, P4, P5, P6, P7, P8, P9
FR-05	Comparison	Cross-project comparison view using historical project data	Display multiple projects side-by-side on the same indicator structure to enable benchmarking and identification of outliers (deviations from prior comparable projects). The benchmark data layer integrates historical project data of the BESS developer (anonymised here as Kappa, Lambda, Mu, Nu, Xi, etc.) as reference for risk indicator calibration and CAPEX/OPEX/revenue assumption checking. (Merged from original FR-05 and DR-03.)	Should	Corroborated	P1, P6, P7, P8
FR-06	Comparison	Standardised assumption list with validation status	List of ~30 key model inputs per project, each tagged as market-generic / region-specific / project-specific, with source attribution (document, person, or estimate) and validation status (well-grounded / partially-grounded / un-grounded).	Should	Corroborated	P6, P8, P9
FR-07	Document processing	Data room document ingestion and structured extraction	Read and parse documents in the data room (contracts, permits, correspondence, technical studies); extract key elements (parties, dates, conditions, restrictions, obligations) into structured form.	Must	Strongly corroborated	P2, P5, P6, P9
FR-08	Document processing	Three-category contract classification (revocability, restrictions, hidden liabilities)	For each contract or right document: (a) classify its revocability as revocable/conditional or irrevocable/unconditional, where irrevocable means the counterparty cannot unilaterally withdraw the right; (b) extract operational restrictions; (c) flag hidden liabilities or cost obligations.	Won't	Role-specific	P2, P9
FR-09	Information completeness	Absence-of-information detection and flagging	For each indicator that should have supporting documentation per project phase, detect when documentation is absent and flag as a distinct status (not low score, not high score: 'unknown / missing').	Should	Strongly corroborated	P1, P2, P4, P5, P6, P9

continued on next page

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
FR-10	Public-source automation	Automated retrieval from public data sources	Programmatic retrieval and structuring of public information per indicator: TSO/DSO grid maps and capacity, Kadaster/cadastre, omgevingsloket/openbaar loket/Bundesland zoning portals, ENTSO-E market depth, meteorological data, KvK/KBO/Handelsregister, sanctions lists, geological/flood/seismic maps.	Must	Strongly corroborated	P1, P2, P3, P5, P6, P7, P8, P9
FR-11	Public-source automation	Cable route geospatial calculation (length $\times$ parcel count)	Given site coordinates and target HV substation, calculate shortest cable route, total length, and number of cadastral parcels traversed. Output as quantitative risk proxy.	Should	Corroborated	P1, P3, P7
FR-12	Public-source automation	Substation expandability assessment	Use public TSO investment plans + aerial imagery to assess substation expandability (reserve fields, topology) for connection viability.	Won't	Role-specific	P3, P7, P9
FR-13	Public-source automation	Permit authority identification and policy retrieval	Given project location, identify the formally bevoegd gezag (issuing authority) and check whether provincial, waterschap, or other approvals are also required per published policy.	Could	Corroborated	P2, P5, P9
FR-14	Monitoring & validation	Continuous public regulatory document monitoring per active jurisdiction	Monitor public regulatory announcements, consultation documents, parliamentary discussions, and regulator publications relevant to BESS per active country/region. Translate findings into projected business-case impact.	Could	Corroborated	P2, P6, P8
FR-15	Monitoring & validation	DD-claim cross-verification against external sources	Where DD documentation contains assertions (e.g. emails claiming municipal support), automatically attempt cross-verification against independent public sources.	Could	Role-specific	P2, P9
FR-16	Monitoring & validation	Business-case-vs-DD assumption reconciliation	Match key line items in the financial model (CAPEX, OPEX, revenue assumptions, timing) against findings in legal/technical/commercial DD reports; flag mismatches and unconfirmed assumptions.	Could	Role-specific	P8, P9
FR-17	Monitoring & validation	Contextual benchmark via cross-municipality energy-transition activity	For permitting risk assessment, retrieve indicators of municipal energy-transition activity (EV chargers, prior renewable permits) as proxy for permit-process speed.	Won't	Singular	P2
FR-18	Output	Rejection-reason logging across projects	Record reasons why projects were not pursued, so institutional knowledge (market entry decisions, counterparty issues, technical disqualifications) is preserved for future reference.	Should	Role-specific	P1, P8
FR-19	Output	Internal pipeline transparency reporting	Surface to the broader organisation which projects are being considered and the reasoning for go/no-go, addressing the ~80% silent rejection rate.	Could	Singular	P3

continued on next page

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
FR-20	Output	Stage-gate cross-functional review support	Structure tool output so it serves as the common input document for a cross-functional (technical/commercial/legal) review before each acquisition decision.	Should	Corroborated	P5, P6, P9

Interface Requirements

Table G.2: Interface requirements.

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
IF-01	Output	Output format compatible with existing reporting flow	Tool output must integrate with existing DD-report Word/PDF format used for the BESS developer's investment proposals to its shareholder.	Could	Role-specific	P1, P3
IF-02	Output	Geographic/map-based primary visualisation	Where applicable, present the project on a map with overlays (HV grid lines, substations, Kadaster boundaries, capacity status) as the primary spatial interface.	Should	Role-specific	P3, P7
IF-03	Input	Connector layer to external public sources	Modular connectors to external data sources (TSO maps, Kadaster, omgevingsloket, ENTSO-E, KvK, weather, geological data) so additional sources can be added without rebuilding the core.	Should	Strongly corroborated	P1, P3, P5, P7, P9
IF-04	Input	Public-source registry with country-specific mappings	Documented registry of public data sources per country (NL, BE, DE), with indicators served by each, update frequency, and access method.	Could	Role-specific	P3, P5, P7
IF-05	Input	Primarily text-based input handling (NLP-orientation)	Most data room content is text (contracts, permits, correspondence, studies). Tool must prioritise document-oriented processing over numerical-modelling capabilities.	Won't	Role-specific	P4, P8, P9

Quality (Non-Functional) Requirements

Table G.3: Quality (Non-Functional) requirements.

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
QL-01	Reliability	Source attribution and per-data-point provenance for every output	Every conclusion, score, indicator value, or extracted fact must be traceable to its source document, public data point, or expert input. No outputs without provenance. This is implemented at the data layer through per-data-point provenance metadata: every data element carries source, retrieval timestamp, and validation status. Both the principle (every output traceable) and the implementation (per-data-point metadata) are non-negotiable. (Merged from original NF-01 and DR-04.)	Must	Strongly corroborated	P1, P2, P3, P4, P5, P7, P8, P9
QL-02	Reliability	Reasoning transparency for computed outputs	For derived scores, predictions, or calculations, expose the underlying logic so a domain expert can validate and override.	Should	Strongly corroborated	P1, P2, P4, P7, P9
QL-03	Reliability	Up-to-date data is non-negotiable	If underlying data (regulatory, market prices, capacity maps) is stale, the tool is not useful. Data freshness is a core reliability criterion.	Should	Role-specific	P6, P8
QL-04	Reliability	Conservative AI hallucination handling	Where AI is used (document interpretation, regulatory extraction), the tool must distinguish 'found' from 'not found' from 'uncertain'. Never invent missing information.	Must	Role-specific	P2, P9
QL-05	Maintainability	Iterative / modular development with rapid feedback loop	Tool should be built incrementally, starting with the most automatable indicators and expanding. Field is moving too fast for a fully specified end-state design.	Could	Singular	P3
QL-06	Security	Confidential / protected processing environment	Documents processed by the tool include sensitive M&A material, contracts, and confidential project information. Must operate in a non-public, access-controlled environment.	Could	Role-specific	P2, P9
QL-07	Adaptability	Country/regional configuration support	NL, BE, DE differ on permitting procedures, grid connection processes, and applicable contract types. Tool must accommodate per-country logic without becoming three separate systems.	Could	Strongly corroborated	P1, P2, P3, P4, P5, P9

Usability/Quality-in-Use Requirements

Table G.4: Usability/Quality-in-Use requirements.

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
US-01	Efficiency-in-use	Helicopter-view as primary output mode	Primary user interaction is a 15–20 minute readable project overview, suitable for non-specialists across functions, with drill-down available for specialists.	Should	Strongly corroborated	P1, P5, P6, P8
US-02	Efficiency-in-use	Time to a first-pass screening record	Should produce a first-pass 152-item record within minutes of machine time and a few hours of reviewer time, measured from the moment the documents are available. External due diligence stays outside this requirement; the intended benefit is earlier risk identification at lower cost.	Should	Role-specific	P1, P2
US-03	Accessibility	Institutional embedding (not person-dependent)	Tool must be usable by multiple people and survive personnel changes. Current experience at the BESS developer: previous tools disappeared when their owner left.	Should	Singular	P3
US-04	Accessibility	Accessibility to non-specialists / multi-disciplinary readability	Output must be readable by people not deeply familiar with every dimension (e.g. Investment Committee members reviewing for FID).	Could	Corroborated	P1, P6, P8
US-05	Output	Long-term revenue forecasting out of scope	20-year revenue forecasts depend on macro and political factors that cannot be reliably automated. Tool flags assumption quality but does not produce its own revenue forecasts.	Won't	Corroborated	P2, P4, P6

Human Factors Requirements

Table G.5: Human Factors requirements.

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
HU-01	Human software interaction	Tool must not replace the financial model	The financial model is complex (6–10 tabs, hundreds of inputs) and must remain in finance team's domain. Tool validates inputs to the model, not replaces it.	Won't	Role-specific	P6, P8
HU-02	Human software interaction	Tool must not replace external advisor sign-off	External advisors' independent sign-off remains required for accountability and bankability. Tool augments DD, never replaces it.	Must	Strongly corroborated	P1, P2, P4, P6, P9

continued on next page

ID	Sub-type	Title	Description	Pri.	Robustness	Sources
HU-03	Human software interaction	Tool must not make autonomous investment decisions	A computer cannot be held accountable. All investment decisions remain with humans; tool produces input to decisions, never decisions themselves.	Must	Strongly corroborated	P1, P2, P6, P8
HU-04	Human software interaction	Mandatory human review layer for all outputs	All tool outputs must pass through human review before influencing decisions. P5: AI misses a share of important issues even with complete documentation.	Must	Strongly corroborated	P2, P5, P6, P8, P9