

From Feasibility to Output Verification

A Task-Level AI Adoption Decision Support Framework
for UX Researchers at Bynder

Author	Zean Wu
Graduation Committee	Evangelos Niforatos - Responsible Supervisor Jorg Van Den Hoven - Supervisor Silvia Ballerini - Company Supervisor
Master Thesis	Ms. Strategic Product Design Faculty of Industrial Design Engineering Delft University of Technology In collaboration with Bynder

Acknowledgements

This project would not have been possible without the people that helped me along the way. I am very grateful for the supportive environment I have and the invaluable feedback I receive throughout the project.

I would like to thank to my university supervisory team, Evangelos Niforatos and Jorg Van Den Hoven, for their continuous encouragement and thought-provoking feedback that made me think critically about my work.

Also, thanks to all of my colleagues at Bynder who took time outside of their work responsibilities to participate in my research sessions. Insights gathered from them really made the project outcomes contextually grounded in the UX research team at Bynder. Special thanks to my company supervisor, Silvia Ballerini, for her continuous guidance and making sure that my project aligned with Bynder's organizational context.

Lastly, thanks to all of my friends and family that offered me encouragement and cheered me on along the way, their emotional support made this project possible.

Abstract

This project investigates how to design a task-level AI adoption decision-support framework that helps UX researchers at Bynder adopt AI responsibly. It responds to a context in which AI adoption is organic and practitioner-led and aims at facilitating the development of standardized governance on AI adoption. The research phase combined a rapid literature scan that identified candidate AI adoption decision constructs, exploratory interviews that surfaced current practice inductively, and construct validation sessions that tested those constructs against UX researchers' own reasoning to confirm or reject those candidate constructs.

These insights were translated into three main interconnecting outcomes: 1. A conceptual reframing of the AI adoption decision from a binary AI use feasibility question into two distinct decisions: whether AI use is feasible, and how human verification of AI output is structured for a given task. 2. A three-layer decision framework operationalizes this through a gating layer, an AI output verification layer that maps task type to suitable verification approaches, and an intensified verification layer that surfaces risk conditions. 3. A prototyped tool, which is grounded in decision support system principles, that translate the framework logic into a tangible tool while preserving researcher judgement rather than automating it.

A formative evaluation engaged UX researchers in think-aloud walkthrough sessions against clarity, usability, and conceptual completeness. Participants found the tool usable, understandable, and coherent and its surfaced factors aligned with their own practice. They also highlighted refinements that were then incorporated into the refined prototype.

Keywords: responsible AI adoption, human-in-the-loop, AI output verification, decision support system

Table of Content

Chapter 1: Introduction	7
1.1 Project Context	8
1.1.1 AI Adoption in Organizational Contexts	8
1.1.2 AI Use in UX Research	9
1.1.3 AI Use in UX Research at Bynder	10
1.2 Project Overview	11
1.3 Project Approach	12
Chapter 2: Theoretical Foundation	13
2.1 Rapid Literature Scan	14
2.1.1 A Three-Level Organizing Structure for AI Adoption Decision Factors	14
2.1.2 Candidate Constructs Identification	15
2.2 Targeted Literature Review	17
2.2.1 Decomposing Task Type: Well-Structured and Ill-Structured Problems	17
2.2.2 Consequence of Error: Dual Roles in AI Adoption Decisions and Evaluation Dimensions	18
2.2.3 Structuring Output Verification	20
2.3 Decision Support System (DSS) Principles	24
Chapter 3: Empirical Foundation	25
3.1 Research Setup	26
3.1.1 Research Approach	26
3.1.2 Participant Profiles	27
3.1.3 Exploratory Interview	28
3.1.4 Construct Validation	29
3.1.5 Data Analysis Process	32
3.2 Research Insights	34
3.2.1 Current AI Use and Challenges at Bynder	34
3.2.2 Construct Validation Results	36
3.2.3 Emergent Constructs Beyond the Original Candidate Construct Set	41
3.2.4 Reframing AI Adoption Decision Process	42
3.2.5 Summary of Empirical Findings	43
Chapter 4: Framework Design	44

4.1 Deriving Design Requirements From Design Implications	45
4.1.1 Consolidation Approach	45
4.1.2 Framework Design Requirements	46
4.1.3 Prototype Design Requirements	48
4.2 Framework Architecture and Decision Logic	49
4.2.1 Overview of the Framework Structure	49
4.2.2 Gating Layer	50
4.2.3 AI Output Verification Layer	51
4.2.4 Intensified Verification Layer	54
4.3 The Prototype in Operation	56
4.3.1 Assessment Flow Walkthrough	56
4.3.2 The Contextualization Feature	65
Chapter 5: Evaluation	67
5.1 Evaluation Goals	68
5.2 Evaluation Approach	69
5.3 Evaluation Results	71
5.3.1 Overall Insights	71
5.3.2 Improvements and Iterations	72
Chapter 6: Discussion	75
6.1 Bridging Literature, Practice and Design	76
6.1.1 Addressing the Research Question	76
6.1.2 Redefining the Theoretical Foundation Contextually	77
6.1.3 Translating Empirical Insights into Design	79
6.1.4 Anchoring Design in Theory	80
6.2 Positioning the Contributions	81
6.3 Limitations & Future Directions	82
6.4 Conclusion	83
6.5 Personal Reflection	84
References	85
Appendix	88

Note on Terminologies

The following terminologies are used with distinct meanings throughout the report:

Constructs and **factors** are used interchangeably. Both terms refer to the specific factors and variables that affect how UX researcher make decisions about AI use for a given task.

The framework refers to the AI adoption decision architecture and logic as the key deliverable of this project.

The prototype (or the prototyped tool) refers to the interactive digital tool that translates the embedded decision architecture and logic into a tangible decision support tool.

A (UX) research task refers to a single unit of work within a larger UX research project at Bynder. It does not refer to the entire project as whole. For example, drafting interview questions and transcribing an interview are research tasks that could belong to the same project.

A User in the context of the framework and the prototype refers to a UX researcher at Bynder, who is the targeted end user of the end deliverables.

1 Introduction

This chapter introduces the background for the project. It begins by positioning the project within the broader landscape of AI adoption in organizational contexts before narrowing down to AI adoption within the specific field of UX research, and finally to the context of Bynder. It then defines the project objective, scope, and the research question the thesis aims to answer. Finally, it provides an overview of the project methodology by detailing the project approach used to move from contextual understanding to framework design.

1.1 Project Context

1.1.1 AI Adoption in Organizational Contexts

The integration of AI into organizations has become a key technological shift organizations face today. However, organizational adoption of new technologies is rarely a top-down and uniform process. Enholm et al. (2022) state that AI is not adopted by organizations through a single, coordinated implementation. Instead, it is adopted gradually and unevenly across teams, functions and individual employees. As a result, organizations usually face challenges in moving from individual experimentation of AI use to a formal and coordinated practice of AI application that generates value.

A common observation across technology adoption in organizations literature is that the early stage adoption of new technology tends to bottom-up. The process is driven by individual employees' experimentations with new tools in their workflows before any organizational policy or shared practice is formalized. This pattern has been documented in the phenomenon of shadow IT, in which employees use unofficially approved tool to complete their work (Klotz et al., 2019). Such bottom-up adoption is not necessarily a failure of organizational governance, but rather a typical early stage of new technology adoption that comes before the development of formal structures and shared norms.

This pattern is also particularly noticeable in the case of gen AI tools due to their easy availability to individual employees when compared to traditional corporate software that typically requires planned organizational procurement and infrastructure building. Bick et al. (2024) confirm this rapid adoption of gen AI trend driven largely by individual employees rather than formal organizational adoption approaches.

The transition from this individually-driven, fragmented model to structuring coordinated and shared approaches represent a common organizational challenge for maturing teams adopting AI across industries. Papagiannidis et al. (2023) states that AI governance is an emergent organization capability that develops in reaction to growing AI presence. Organizations need to build practices that reduce AI adoption risks while preserving the value individual employees are already creating with the use of AI.

1.1.2 AI Use in UX Research

UX researchers are increasingly adopting AI tools in their workflows. AI is being embedded across different stages within the UX research workflow. A recent industry study surveyed 485 UX researchers and found that 80% now use AI in their work, which is a 24 percent increase from the past year (Steelman & Koufos, 2025). Moran and Rosala (2024) state that UX researchers use AI to support tasks spanning across multiple stages of the UX research process, from project planning and generating interview questions to transcribing interview sessions and qualitative data analysis. What’s more, Williams (2024) states that AI technologies are being used increasingly in tasks that have been traditionally carried out completely by human researchers such as data coding and theme identification. The outputs of such tasks would inform downstream product decisions.

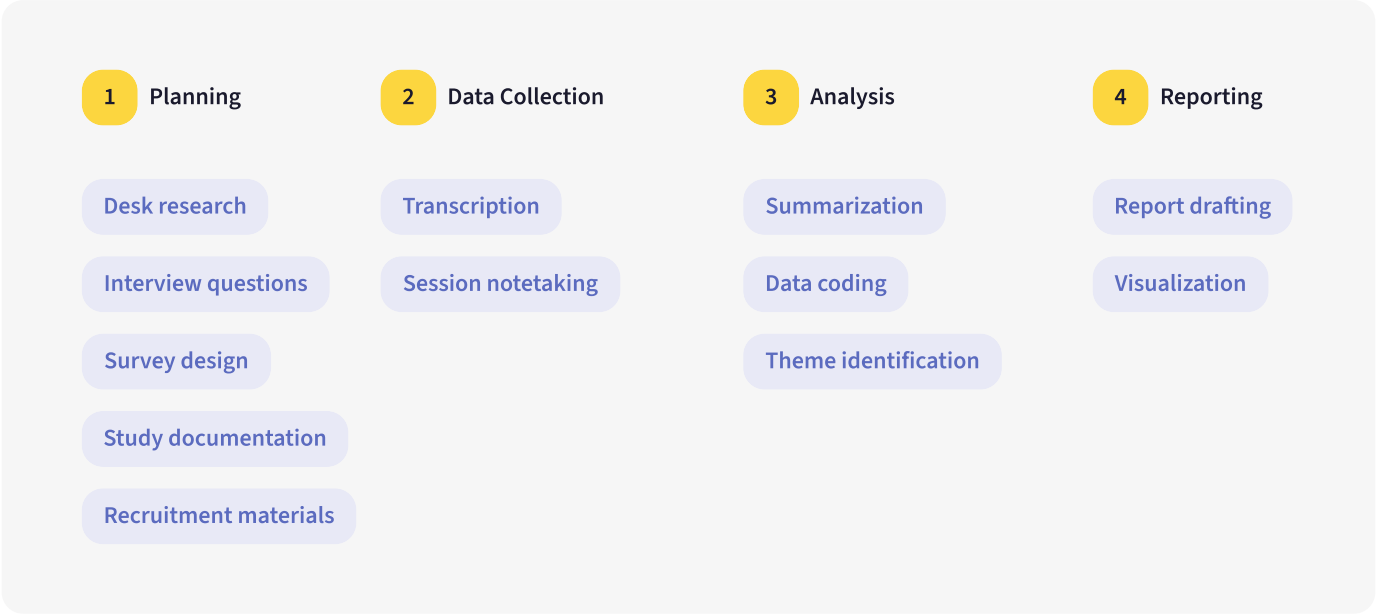


Figure 1: examples of UX research tasks where AI tools are currently being used, Moran and Rosala (2024)

1.1.3 AI Use in UX Research at Bynder

Bynder is an enterprise software company that provides cloud-based digital asset management (DAM) solutions. It provides a centralized, AI-powered system for organizing, storing, and sharing brand assets like logos, images, and videos (Bynder, n.d.). Within Bynder, a small team of UX researchers as part of the UX team supports product development by conducting UX research activities such as user interviews and usability testing. The research outputs will then inform design decisions, product roadmaps and organizational priorities. UX designers, product managers and leadership are stakeholders who rely on these research outputs to make informed decisions about what to build and prioritize.

Consistent with the broader trend of increasing adoption of AI in the general UX research field described in Section 1.1.2, AI tools have also been increasingly adopted in UX researchers' workflows at Bynder. During prior foundational observations within the product team, the author has observed that UX researchers were using AI for tasks such as transcribing interview sessions, clustering themes and synthesizing interview insights. As mentioned in Section 1.1.1, organizations across industries commonly go through an early stage of technology adoption in which the use is more bottom-up and driven by individual employees. The current state of AI adoption in UX research at Bynder is a local case of this broader organizational technology adoption pattern, which allows the team to discover where AI adds the most immediate value.

With this initial phase of organic, bottom-up experimentation of AI adoption, high-impact use cases of AI have been successfully identified. The natural next step is for Bynder to transition into more standardized governance on AI adoption. As a result, there is an opportunity to facilitate this transition by building a shared logic behind AI adoption. While existing work tends to address organizational AI adoption broadly, there is a notable absence of task-level AI adoption decision framework specifically for UX researchers. This project aims to develop a structured AI decision framework that makes the decision logic behind AI use explicit and consistent across the UX research team instead of leaving AI adoption decisions fully to varied individual judgements. The framework would support UX researchers in making informed decisions when it comes to using AI for a given research task while ensuring that their own professional judgement is still preserved.

1.2 Project Overview

UX research is a meaningful domain for creating an AI adoption decision framework since AI adoption in the field is rapid, and UX research outputs would feed directly into downstream design and product decisions, which makes it a high-stakes area for exploring AI adoption decision making. The primary objective of this thesis project is to conceptualize, design, and validate a structured AI adoption decision framework for UX researchers at Bynder. This project aims to address this by creating a shared, clear decision logic that helps UX researchers develop a responsible and consistent approach to using AI for a given research task. Importantly, a prototyped tool is designed to operationalize the framework for decision support on AI use rather than decision automation. It aims at structuring and making the decision process of using AI more transparent and responsible while ensuring the professional judgement of UX researchers are preserved.

The scope of this project is defined by the following parameters:

- **Primary Users:** The framework and the prototyped tool are intended to be used by UX researchers at Bynder. Although the underlying decision logic behind the framework could hold true for employees in other roles, all empirical fieldwork is scoped to UX researchers.
- **Organizational Context:** The findings from fieldwork and end deliverables are restricted to the specific context of Bynder. While a certain degree of universal generalizability of the framework could exist, proving it lies outside of the scope of this project.
- **AI Use Decision Focus:** The framework and the accompanying prototyped tool aim at helping UX researchers develop a comprehensive AI engagement strategy for a given research task. They are not intended to recommend or evaluate specific AI softwares.

To achieve the project objective, the study will be guided by the following research questions:

Main Research Question

How should task-level AI adoption decision for UX researchers at Bynder be structured to ensure responsible practice?

Sub Research Question 1

How do UX researchers at Bynder currently make AI adoption decisions, and what are the potential opportunities in this era where AI adoption is organic and practitioner-led?

Sub Research Question 2

What task, human, and organizational factors should govern a structured approach to AI adoption decisions for UX researchers at Bynder?

Sub Research Question 3

How should the identified AI adoption decision factors be operationalized into a framework architecture and a prototyped tool for UX researchers at Bynder?

1.3 Project Approach

This project applies an integrated design research approach of research-to-design and research-through-design, which means that theory and fieldwork don't happen in sequence, but rather continuously inform each other throughout. The project is divided into three phases:

Phase 1: Understand

The first phase starts with a rapid literature scan across five domains that cover the three level organizing structure for AI adoption decision (task, human, and organizational levels). These domains are cognitive task analysis, AI capabilities in qualitative and UX research, human oversight of AI systems, trust in automation, and AI governance. The literature scan is done broadly rather to compile a set of theoretically grounded candidate constructs relevant to AI adoption decisions. These candidate constructs are then brought into research sessions with UX researchers at Bynder where participants are first encouraged to openly share their experiences with AI use before being asked to validate these theoretical constructs. This activity functions as an empirical filter to ensure constructs of the proposed framework are confirmed based on practitioner judgment rather than theoretical assumption alone. The confirmed constructs that emerge will be the foundation for targeted literature review to develop the consolidated design requirements for the proposed framework.

Phase 2: Build

The second phase translates the validated constructs and design requirements generated in Phase 1 into the AI adoption decision framework that guides UX researchers in determining whether AI use is permissible for a given research task and how AI output verification should be structured. In parallel, informed by decision support system (DSS) literature, a prototype decision support tool is developed to operationalize the framework.

Phase 3: Validate

The third phase validates the framework and prototype. An evaluation session is conducted with UX researchers at Bynder to assess the framework logic embedded in the prototype and interactions of the prototype. Findings from the evaluation sessions inform design iterations.

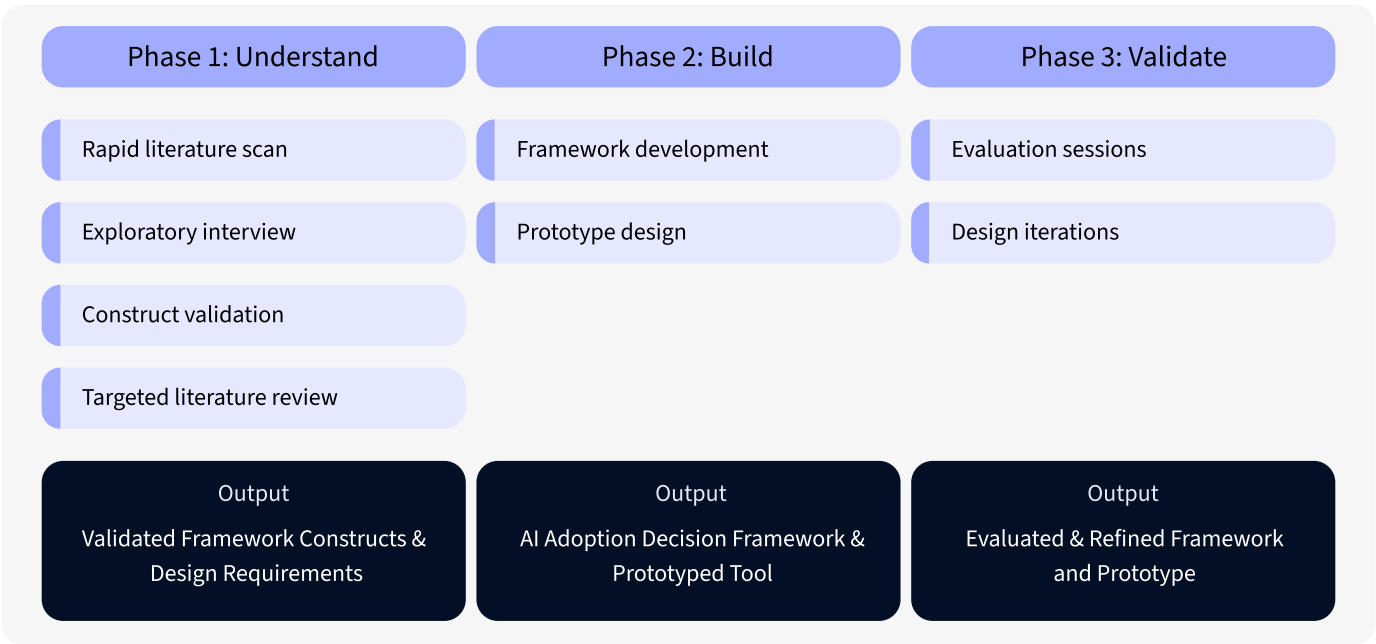


Figure 1: project approach overview

2 Theoretical Foundation

This chapter presents the theoretical foundation that supports the AI Adoption Decision Framework. The literature review is conducted in two stages. First, a rapid literature scan across five domains is conducted to identify a set of theoretically grounded candidate constructs relevant to AI adoption decisions (Section 2.1). After empirical validation of these candidate constructs with UX researchers, a targeted literature review is conducted to answer the specific theoretical questions that the empirical findings surfaced (Section 2.2). Therefore, this section is read as a response to Chapter 3. In the meantime, decision support system (DSS) literature is reviewed to inform the design of the prototype (Section 2.3).

2.1 Rapid Literature Scan

2.1.1 A Three-level Organizing Structure For AI Adoption Decision Factors

When UX researchers make decision on how AI should be adopted, specifically when deciding on the appropriate degree of AI involvement and human oversight for a given research task, the process is never straightforward and the decision cannot be reduced to a single dimension.

Existing literature establishes the notion that technology adoption is multi-dimensional and provides foundations for distinguishing these dimensions. Parasuraman et al. (2000) argue that automation decisions should begin with what type of work is being automated and consider task characteristics as a primary input. Meanwhile, Lee and See (2004) show that trust plays a role in determining the relationship between automation capability and actual reliance, and that this trust varies across individuals. This insight establishes the human operator as an independent dimension when shaping how AI is used. Finally, research on AI governance (Floridi et al., 2018) shows that organizational factors, such as organizational policies and accountability structures, serve as boundary conditions on adoption regardless of task characteristics or individual differences.

💡 How does it inform my project?

After synthesizing insights from existing literature, this study adopts a three-level organizing structure for the AI Adoption Decision factors: Task level, human level, and organizational level. The task level addresses factors that are relevant to the property of the given research task . The human level addresses factors that are relevant to the researcher's trust in AI and personal judgement on AI use. The organizational level addresses the institutional factors within the organization that govern AI use. This structure is not borrowed from any single existing framework but represents a synthesis across the AI automation, human factors, and AI governance literatures. It provides the organizing structure both for AI adoption decision factors during the construct validation process with UX researchers (Chapter 3) and the framework (Chapter 4).

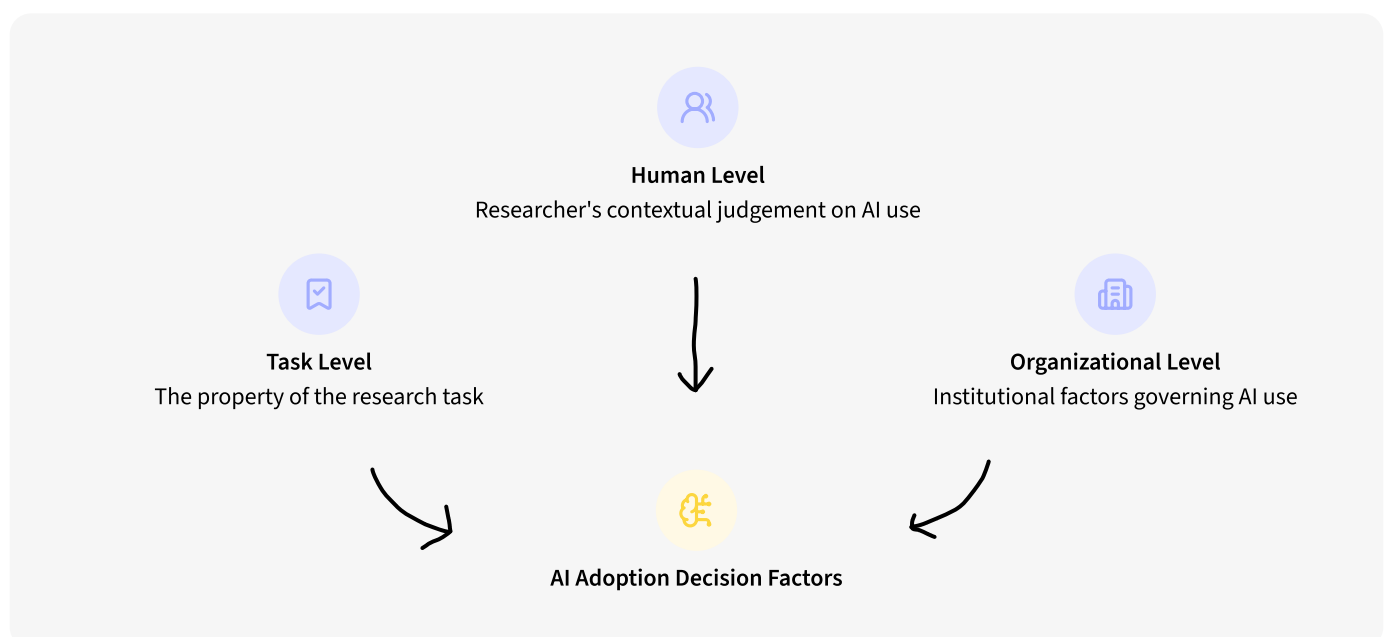


Figure 2: three-level of organizing structure for AI adoption decision factors

2.1.2 Candidate Constructs Identification

Upon establishing the three-level organizing structure, the following step was to identify a set of candidate constructs for the AI adoption decision framework. These constructs refer to the specific factors within each level that affect how UX researchers decide to use AI in a given task within the UX research workflow. (Note: throughout this thesis, the terms 'construct' and 'factor' are used interchangeably). A rapid literature scan was conducted across five relevant domains mapped to the three levels of the organizing structure: cognitive task analysis, AI capabilities in qualitative and UX research, human oversight of AI systems, trust in automation, and AI governance. The goal was to identify theoretically grounded candidate constructs relevant to AI adoption decisions for the construct validation sessions with UX researchers.

Task level - cognitive task analysis literature was reviewed to substantiate that the cognitive demands of a task are relevant to how suitable it is for AI automation. Also, literature on AI capabilities in UX research was reviewed to prove that AI capabilities varies across research task types and thus should be considered a factor in AI adoption decision. Two candidate constructs were derived from the relevant literature.

- **Construct 1: Task Cognitive Demand**, which refers to the amount of active reasoning and interpretation a task requires rather than following clear rules or procedures. Rasmussen's (1983) skill-rule-knowledge framework proves that tasks differ in their cognitive demands. The corresponding cognitive demands can be categorized into three groups: automatic, skill-based behavior to knowledge-based reasoning through unfamiliar situations. This variation in cognitive demand is highly relevant to the task's suitability to AI automation.
- **Construct 2: Current AI Capability**, which refers to the realistic performance of current AI tools on given research tasks. Hitch (2024) establishes that AI performance varies greatly on different task types. For example, current AI performs well on structured tasks like summarization and it is not as proficient in interpretive tasks like thematic analysis, which shows that AI capability is not uniform and should be assessed based on the given research task.
- **Construct 3: Output Verifiability**, which refers to whether the UX researcher can check the correctness of AI-generated output. Some task outputs are easier to verify than others. For example, an interview transcript can be easily verified against the original audio while outputs that require interpretation, such as thematic coding and persona generation, are harder to verify as there is no single correct answer. Langer et al. (2024) state that the ability to detect errors in AI output is a prerequisite for effective human oversight, and that this ability varies depending on the inherent quality of the task and its output.

Human level - literature on trust in automation (Lee & See, 2004) shows that individuals trust automated systems at different degrees and thus shape how they engage with them. Trust itself is not a standalone construct but rather an umbrella term that encompasses multiple human-level factors. Literature on trust in automation informed the identification of two key candidate constructs on the human level.

- **Construct 4: Domain Expertise**, which refers to how experienced the UX researcher is with the given research task. Parasuraman and Riley (1997) argue that an operator's level of expertise affects their vulnerability to errors in automation, which indicate that domain expertise is a relevant factor in determining AI adoption since individuals' ability in detecting errors in AI use for a given task differ depending on their past experiences with similar tasks.
- **Construct 5: AI Literacy**, which refers to the skill level of the UX researcher in effective AI use. This includes ability to craft effective prompts and understand AI tool limitations. Christou (2024) states that the effectiveness of AI use in qualitative analysis is dependent on the user's ability to prompt iteratively and

understand what AI can and cannot do. Therefore, AI literacy is a relevant factor on the human level for the AI adoption decision.

Organizational level - literature on AI governance was reviewed to gather organizational factors relevant to decisions on AI use regardless of the inherent nature of the given task or individual differences. Three candidate constructs were derived from relevant literature.

- **Construct 6: Transparency Expectation**, which refers to whether the downstream stakeholders receiving the research results expect to know how AI was involved. Floridi et al. (2018) identify explicability as an essential principle for responsible AI use. What's more, Jobin et al. (2019) concludes that transparency is the most frequently cited principle across frameworks through a review of 84 AI ethics guidelines, which further proves that transparency expectation is a valid construct on the organizational level for the AI adoption decision.
- **Construct 7: Consequence of Error**, which refers to the amount of damage an error in the research output would cause. The same research task where AI is used can have different levels of consequences depending on the context in which the research output is used. Parasuraman and Riley (1997) supports it by stating that the consequences of automation failure vary across contexts.

💡 How does it inform my project?

The rapid literature scan identified seven candidate constructs across the three levels of the organizing structure: cognitive demand, AI capability, and output verifiability at the task level; domain expertise and AI literacy at the human level; and transparency expectation and consequence of error at the organizational level. They were brought into the construct validation session with UX researchers (Chapter 3). During these sessions, UX researchers evaluated what constructs they find the most relevant to their decision making process when it comes to AI use. Confirmed constructs were carried forward as foundations for the framework design.

2.2 Targeted Literature Review

2.2.1 Decomposing Task Type: Well-Structured and Ill-Structured Problems

In the construct validation study (Chapter 3), participants identified task cognitive demand as one of the three most prioritized AI adoption decision factors. However, the result also revealed a limitation in the original definition of task cognitive demand as a construct. As stated in Section 2.1.2, Rasmussen (1983) defined task cognitive demand as a single spectrum from low to high. While this definition captures the intensity difference in cognitive tasks, participants described that two tasks with similar cognitive demand could adopt entirely different AI adoption strategies, depending on the task type. This section reviews literature on cognitive task typologies to distinguish different types of cognitive work and synthesize knowledge on how they would inform the AI adoption strategies differently.

Jonassen (2000) provides the theoretical foundation for the distinction of task types. His problem typology categorizes tasks by their structural properties, which include the nature of their solution space, the clarity of their parameters, and the criteria used for evaluating solutions.

Well-structured problems have limited parameters, clearly defined rules that guide the solution process, and solutions that can be evaluated against established correctness criteria (Jonassen, 2000). In UX research at Bynder, examples include synthesizing interview data into thematic clusters against predefined categories, reviewing and summarizing existing knowledge across various documents, and compiling customer lists according to defined sampling criteria. These tasks are cognitively demanding, but their output can be assessed against existing standards.

Ill-structured problems have multiple potential solutions, fewer adjustable parameters, and uncertainty about which concepts or principles are needed for a solution (Jonassen, 2000). According to Jonassen (2000), design problems are among the most ill-structured since they often involve vaguely defined goals and require the solver to defend their subjective judgments. In UX research at Bynder, examples include defining research objectives, formulating UX recommendations and defining product directions (tasks that require forward-looking strategic judgment) are tasks that share these properties.

This distinction directly explains why tasks with similar cognitive demand level require different AI adoption strategies. Cropley (2006) also argues that cognitive processes of tasks differ qualitatively, not only in intensity. Well-structured tasks have known criteria, so AI can take on a greater execution load and verification can be procedural. In comparison, ill-structured tasks have no single correct answer. Although AI can assist in exploring solutions, the final verification requires professional judgment. Therefore, AI adoption strategy needs to adjust depending on the task's structural properties.

Design Implication

Jonassen's task typology and Cropley's evidence on cognitive science establish that tasks differ not only in the intensity of required cognitive effort but also in the type of cognitive work involved, which is also what UX researchers at Bynder prioritized when it comes to making AI adoption decisions. This distinction provides the theoretical basis for decomposing the task cognitive demand construct by type as the operative dimension of the framework in the framework design (Chapter 4).

2.2.2 Consequence of Error: Dual Roles in AI Adoption Decisions and Evaluation Dimensions

The result from the construct validation activity (Chapter 3) identified consequence of error as one of the top three most prioritized AI adoption decision factors. What more's, it revealed two potential dual roles of this factor. One participant treated it as a factor that caps AI involvement when it is extreme enough. In comparison, another participant treated it as a modifier of verification intensity of AI output and did not consider it as a factor that limits AI use in the first place. The empirical data does not provide a theoretical basis for whether both roles coexist, nor does it surface dimensions regarding how consequence of error should be evaluated. As a result, this section reviews literature to justify the dual-role of consequence of error and identify dimensions for evaluating consequence of error.

Consequence of Error as a Cap on AI Involvement

Bigman and Gray (2018) state that stakes severity changes the extent to which users consider AI appropriate for a task. For example, in situations that involve moral judgement or life-or-death decisions, users refuse to delegate tasks to AI entirely, regardless of AI's accuracy in principle. Steyvers and Kumar (2024) also confirmed this qualitative shift in how user apply AI when stakes are severe.

In Bynder's B2B SaaS business context, consequence of error of AI does not reach the critical human welfare level where strict refusal of AI use would apply. Within Bynder, it serves as a cap on AI involvement. According to Parasuraman et al.(2000) taxonomy of automation levels, automation can operate at different levels, depending on a set of criteria that includes consequences of failure. The literature justifies the existence of capping when consequence of error is high. What this taxonomy model leaves open is the exact threshold of capping, which is set by the UX researcher themselves according to their own judgements on the task context. The form, meaning how much the threshold of AI involvement is lowered and in what ways, that capping takes is also not defined intentionally. This framing aligns with UX researcher's first framing of consequence of error as a cap on AI involvement but intentionally not specifying what form capping takes (Section 3.2.2).

Consequence of error as a modifier of Verification Intensity

Research has also shown that higher consequence of error or stakes lead to more careful engagement with AI output rather than directly determining whether AI is used at all. Langer et al. (2024) make the statement that when consequences or error are high, overseers increase the rate of scrutiny applied to AI output. Such behavior operates within continued use of AI rather than triggering AI avoidance.

Similarly, Hoff and Bashir (2015) identify risk as a primary situational factor that modifies how users engage with automated systems. Higher perceived stakes change scrutiny behavior without eliminating dependence on automated systems. Therefore, these findings support UX researcher's second framing that when consequence of error is high, they verify AI output more intensively but they do not stop using AI (Section 3.2.2).

Evaluation Dimension for Consequence of Error

Literature has also shown that consequence of error is a multi-dimensional construct. Two dimensions emerged after reviewing relevant literature. Each dimension captures a distinctive aspect of how consequence of error affects AI adoption decision.

The first dimension is **severity**. It refers to the degree of harm an error causes and the number of parties affected. Severity is a consistent dimension mentioned across literatures on AI and automation (Hoff & Bashir, 2015). Slovic's (1987) psychometric paradigm of risk perception states that the amount

of people or party exposed to a risk affects its perceived severity, which indicates that scope as a factor should also be included within the severity dimension rather than being a separate factor.

The second dimension is **reversibility**. It refers to the degree at which an occurred error can be corrected or undone. Slovic (1987) proves the existence of this dimension on their research on risk perception, which shows that people perceive risks as more severe when potential outcomes are uncontrollable or even catastrophic.

Design Implication

First, the framework should treat consequence of error as a dual-role construct: either capable of capping AI involvement or as a modifier of verification intensity. It does not define a universal cutoff value for when capping happens nor specify the form (how much AI use is reduced and in what ways) that capping takes. These are determined by the researcher's own judgement and responsibility contextual to the project and Bynder's organizational stakes. Second, the framework should assess consequence of error around two dimensions: severity and reversibility. The framework's role is to provide guidance by structuring the researcher's own reasoning on AI adoption decisions.

2.2.3 Structuring Output Verification

The result from the construct validation activity (Chapter 3) show that output verifiability is the highest-priority construct (75 percent of participants ranked it as a top three construct, deal-breaker status from two participants). However, the studies also revealed that UX researchers lack structured AI output verification approaches. This section reviews literature on verification practices in qualitative research and human oversight of AI systems to set up a theoretical foundation for structuring AI output verification. This section concludes with four verification methods that would inform the framework design (Chapter 4).

Conditions for Effective Error Detection in AI Output

Langer et al. (2024) identified two factors that determine whether someone can effectively detect errors in AI output. The first factor is sensitivity, which refers to the ability to distinguish correct outputs from incorrect ones. The second factor is response bias, which refers to the threshold at which the person acts upon perceived evidence of error.

In the context of AI output verification, there are three conditions that shape sensitivity. The first factor is the availability of ground truth, meaning that when predefined assessment criteria or original source exist, the person who oversees the AI output can then compare AI output against them (Langer et al., 2024). This condition is linked to the task typology in Section 2.2. Well-structured tasks usually have available ground truth, while ill-structure tasks don't. Secondly, an individual's domain expertise also improve their ability to detect errors because they could reference internal standards on what quality output looks like (Parasuraman & Riley, 1997). Lastly, the intrinsic property of large language models, that they can present fabricated information with the same confidence level as accurate content, reduces sensitivity. Plausible information generated by AI does not signal correctness, and active verification is required.

Regarding response bias, it is shaped by contextual factors. Langer et al. (2024) state that when the consequence of error is high, the person checking the AI output tends to flag potential errors more proactively and require less evidence before flagging them. This indicated that the consequence of error may also modulate verification intensity and how individuals act upon perceived evidence of error in AI outputs.

In addition to sensitivity and response bias, Bucinca et al. (2021) state that people have the tendency to systemically over rely on AI output, even when it is incorrect. Encouraging independent judgement and analytical processing before seeing AI output can reduce people's over reliance on AI output. This suggests that the structure of the verification process also impacts effective error detection in AI output.

AI Output Verification Approaches

In the field of qualitative research, output verification strategies have already been established under the concept of trustworthiness, Korstjens and Moser (2018) state that strategies include triangulation, which refers to cross-referencing across multiple sources or methods; audit trails, which refers to creating a documented record of how decisions were made and methodological choices throughout the research process so that the research could trace how raw data was transformed into conclusions; and member checking, which refers to a practice in which the researcher returns the findings to the original research participants to check whether the analysis actually represent their actual perspectives. However, these strategies need to be adapted in the context of AI-generated output. When AI generates the output, the researcher does not conduct the analysis of source data themselves. Instead, they review the end output.

Making use of the literature on verification strategies for qualitative research output and human oversight on AI output (Section 2.3.1), four AI output verification approaches are identified. Each one of these approaches operate in a distinctive task context.

Approach 1: Source Tracing

Source tracing refers to checking AI-generated output against the original source materials by locating specific data points such as document excerpts and interview quotes to evaluate whether the AI output genuinely represent them.

This approach adapts the audit trails strategy for AI contexts. The audit trail documents the researcher's own reasoning of key decision points from original data to conclusions. In AI contexts, AI does the reasoning but the reasoning process is not generally transparent. Therefore, the audit trails become source traceability in the context of AI-generated output to make sense of AI's reasoning process. The same principle of traceable evidence in audit trails still applies in verifying AI output. Source tracing also agrees with Langer et al.'s (2024) argument that when ground truth is available, error detection sensitivity increases.

Therefore, the source tracing approach is most applicable to well-structured tasks where the AI output can be traced directly back to the source data. However, it can also be applied to any factual element embedded within the AI output of ill-structured tasks (ex. checking the empirical evidence of a UX recommendation) but it does not verify the interpretive element itself (ex. the specific UX recommendation), which requires the verification approaches discussed below.

Design Implication

The framework should position source tracing as a key verification approach for well-structured tasks with accessible source data.

Approach 2: Criteria-Based Checking

Criteria-based checking refers to evaluating AI output against predefined evaluation criteria, where those criteria specify objectively verifiable qualities of what acceptable outputs should look like.

Criteria-based checking relies on the predefined-criteria condition as a form of ground truth (Langer et al., 2024). Examples of criteria-based checking include checking whether a thematic analysis covers the expected range of topics, or whether a data synthesis addresses the required scope and predefined research questions.

Design Implication

The framework should position criteria-based checking as a verification approach for tasks that include predefined criteria.

Approach 3: Comparative Verification

Comparative verification refers to the process of checking AI output against an independent knowledge source to surface inconsistencies.

Examples of independent knowledge sources include outputs from different AI tools, and relevant internal or external documents.

This approach adapts the triangulation strategy for AI contexts. Triangulation cross-references across multiple sources or methods to further provide evidence on research output in traditional qualitative research (Korstjens and Moser, 2018). For AI output, since the AI internal reasoning is opaque, comparative verification becomes especially important.

Comparative verification can be applied across task types, but its role differs by task type. For well-structured tasks, it serves as a supplementary check. It's mainly applicable to ill-structured tasks in which objectively correct criteria do not exist and thus comparison with another independent knowledge source becomes a way to verify AI output.

Design Implication

The framework should position comparative verification as a verification approach mainly for ill-structured tasks.

Approach 4: Independent Judgement Before Using AI

Independent judgement before using AI requires the researcher to articulate a preliminary interpretation of what the data should show before exposure to AI output, and then use those pre-AI interpretation to compare against the AI output.

Examples of independent judgement before using AI includes noting down the dominant themes that the research already sensed during the interviews they conducted and drafting a brief, initial interpretation of the research data. According to Buçinca et al.'s (2021)'s cognitive forcing research, requiring analytical engagement before exposure to AI would prevent anchoring, which in this context refers to the researcher's interpretation being shaped by AI output before they've formed their own. Preventing anchoring would then preserve the independent stance needed to detect errors in AI output.

While this approach can be applied to both task types, it is a less suitable and efficient approach for well-structured tasks though the mechanism would still work. This is because for well-structured tasks, source tracing and criteria-based checking can be used to check external ground truth to directly address the correctness of the AI output.

Design Implication

The framework should position independent judgement before using AI as a starting point for ill-structured tasks.

Verification Approach	Purpose	Failure It Prevents
Source Tracing	Grounding Check	AI Hallucination
Criteria-Based Checking	Quality Assurance	Application errors (ex. out-of-scope Output, inclusion errors)
Comparative Verification	Triangulation (Convergence Check)	Single source anchoring of AI output
Independent Judgement Before Using AI	Cognitive forcing on personal analytical engagement	Over-reliance on AI output

Figure 3: AI output verification approaches summary

2.3 Decision Support System (DSS) Principles

The previous sections establish the theoretical foundation for the framework. In order to operationalize the proposed framework in practice, a prototyped tool was also designed as a deliverable of the project. This section reviews literature on decision support systems (DSS) to generate high-level design principles for the prototype.

According to Power (2002), DSS is defined as an interactive system designed to support humans in making decisions for problems that are not well-structured by combining analytical models with user input. In the case of the prototyped tool, the proposed decision logic for the framework would function as the analytical model while the UX researcher's input would function as user input. It is important to draw the distinction between decision support and decision automation for the proposed prototype. Decision automation systems (DAS) make the final decision and replace human judgements with algorithmic output. In contrast, the proposed prototype is a DSS that preserves human agency, which means that it focuses on supporting and improving human decision-making rather than automating it (Arnott & Pervan, 2014).

DSS literature has identified several design principles that informed the prototype design. The first key principle is **Reasoning Transparency**. A DSS should make its decision logic transparent to the user rather than presenting recommendations as a black box. Cummings (2006) states that when users do not have visibility into how a recommendation was generated, they are more likely to either blindly follow it or reject it completely, which defeats the purpose of decision support.

Design Implication

The prototyped tool should make visible how the outputs are derived from the user inputs to ensure reasoning transparency.

The second principle is **Mitigation of Automation Bias**. According to Goddard et al. (2012), a systematic review has shown that users tend to over rely on DSS output, particularly when recommendations are presented in a prescriptive fashion. This effect is particularly stronger when users are less confident in their own judgment. The goal of having this principle is to prevent automation bias and “autopilot” mode of the user.

Design Implication

The user's independent judgement should be encouraged.

The third principle is **Preservation of User Control**. Silver (1990) states that less restrictive designs maintain the user's autonomy to use their judgment instead of directing them toward a specific outcome.

Design Implication

The tool should present its output in a suggestive way and be used as a starting point that the user can modify, override, or disregard rather than as a fixed prescription.

3 Empirical Foundation

This chapter contains the empirical foundation of the project. It presents the research sessions conducted with UX researchers to ground the framework in practical UX research workflow at Bynder. The sessions are made up of two parts: exploratory interviews to understand current AI use in the UX research workflow at Bynder, followed by a construct validation session in which UX researchers evaluated the seven candidate constructs (Chapter 2). The chapter concludes with insights that were carried forward to inform the framework design (Chapter 4).

3.1 Research Setup

3.1.1 Research Approach

The seven candidate constructs identified (Chapter 2) were derived from relevant academic literature. Although they are theoretically established, they have not been evaluated by the end users of the proposed framework. The goal of this empirical research phase is to bridge this practicality gap by bringing practitioner inputs into the framework constructs identification process before proceeding to the design phase.

To achieve this goal, UX researchers at Bynder were invited to research sessions (each session lasted for 45 to 60 minutes approximately). Each session is composed of two parts in an intentional sequence: a semi-structured, exploratory interview as the first part and a construct validation activity as the the second part. This specific sequence adopted an inductive-deductive research approach logic. The interview is inductive in nature where open-ended questions would allow UX researchers to openly share their own experiences with AI use in their workflow and decision making process without being exposed to theoretical constructs. After that, the construct validation activity takes a deductive approach where UX researchers are presented with theoretical constructs on AI use decision making to evaluate against their own experiences. Fereday and Muir-Cochrane (2006) states that combining inductive and deductive approaches improves qualitative research quality by allowing empirical patterns and theoretical frameworks to inform each other. The intentional sequencing of inductive first, deductive second allows genuine comparison between what factors UX researchers actually consider when making decisions on AI use and what the literature identified without being biased by theoretical constructs.

Participants profiles are UX researchers at Bynder with varied seniorities, ranging from junior to senior levels. They were selected through convenience sampling given the small team size and the project’s organizational scope. All participants are UX research practitioners who actively engage in research tasks in their positions. Therefore, they directly represent the intended user of the proposed framework.

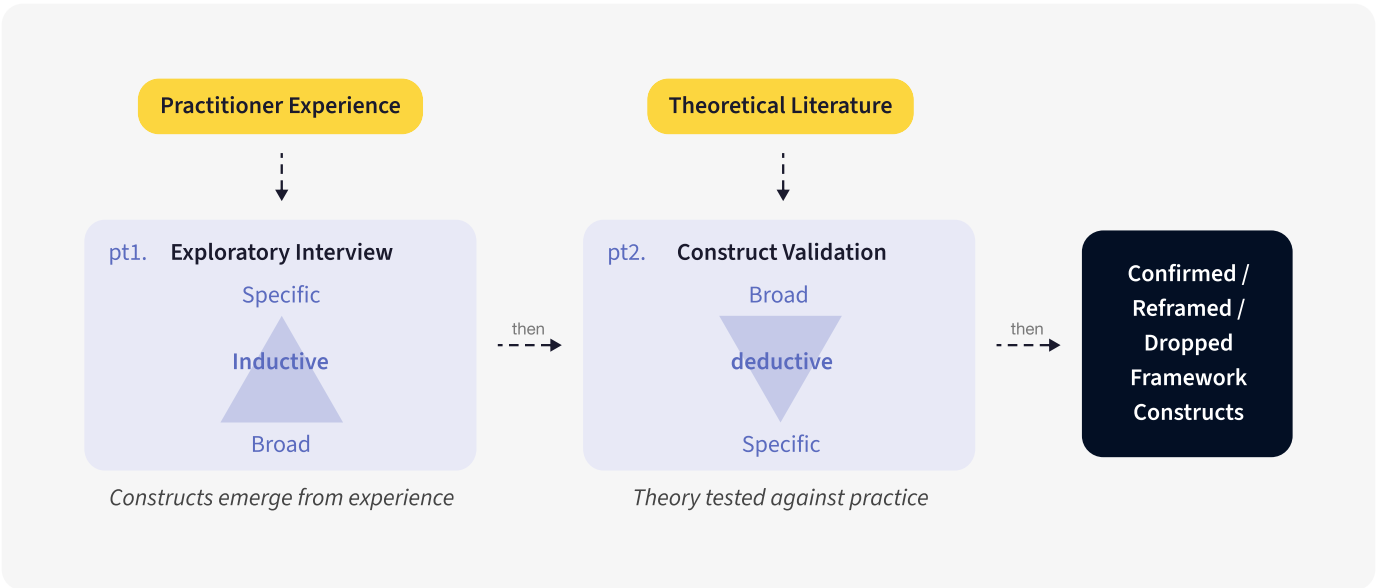


Figure 4: research approach overview

3.1.2 Participant Profiles

A representative sample of UX researchers from Bynder’s UX research team participated in the research sessions. Participants were chosen through convenience sampling given the project’s organizational scope within Bynder. All participants are active UX research practitioners who regularly conduct UX research tasks. These tasks generally fall under the following categories:

- 1. Discovery & Understanding (ex. User interviews)
- 2. Evaluation & Testing (ex. Usability testing)
- 3. Quantitative & Analytics (ex. Analyzing usage analytics)
- 4. Synthesis & Sense-making (ex. Writing research reports and recommendations)
- 5. Ongoing (ex. Maintaining a research repository)

Attribute	Description
Sampling	A representative sample of UX researchers from Bynder’s UX research team
Professional Experience	Mid-level to senior-level industry experience in user research
Prior AI Usage	All participants had prior experience using AI tools in their workflows though the extent and sophistication vary



Figure 5: participant profiles

3.1.3 Exploratory Interview

The first part of each session was a semi-structured exploratory interview aimed at surfacing how UX researchers currently make decisions about AI use in their workflow before introducing the theoretical constructs (chapter 2) for evaluation. The interview addressed three key areas with the following interview questions.

The first area addressed was current AI use. It was guided by the following research question:

- Can you give me a rundown of how you currently use AI in your workflow? Which tools and for which tasks?
- When you've used AI for a task, what do you do with the output before using it? (If the participant gives generic answer like “I review it.”, then a follow-up probe “Can you walk me through what the review process looks like?” will be asked.

💡 How does it inform my project?

The questions helped establish a basic understanding of how AI is currently incorporated into the UX research workflow.

The second area addressed was challenges around AI use. It was guided by the following research questions:

- What was your biggest challenge when using AI in your research?
- Were there tasks where you've tried using AI and stopped using it? Or changed how you use it? If so, why?

💡 How does it inform my project?

The questions helped understand what makes UX researchers hesitant when using AI and where uncertainty arises, which provided insights into practical challenges UX researchers encounter when using AI. These insights in turn contextualized the problem space in which the proposed framework operates.

The third area directly addressed decision factors. It was guided by the following research questions:

- How do you decide whether and how to involve AI in a research task?

💡 How does it inform my project?

The question addressed practical decision factors. This part was inductive in nature since it allowed UX researchers to freely articulate their own decision criteria before presenting them the derived constructs from literature, which enabled comparison between factors that emerge from practical UX research practices and those identified in literature.

All interviews were conducted individually to ensure each participant's perspective was independently captured, rather than being potentially affected by group think. Sessions were recorded and transcribed for further analysis with participant consent.

3.1.4 Construct Validation

The second part of each session was the construct validation session in which the seven candidate constructs identified (Chapter 2) were presented to UX researchers for validation. Each construct was presented as an individual card with its name, an illustration and its definitions in plain language. Concrete examples of each construct grounded in UX research tasks were also provided to the participants to ensure correct understanding. The validation activity was structured around three areas of inquiry to assess **relevance**, **behavior and relationships**, and **completeness** of the candidate constructs.

In the relevance assessment stage, participants were asked the following questions:

- Is this (the construct) something you consider or think you should consider when making decisions about AI use for a research task? Can you provide one example?
- If you had to pick the three most important ones, which would they be?

💡 How does it inform my project?

The first question confirmed whether each construct is relevant to UX researchers' actual decision making process for AI use grounded in practitioner judgement. The second question informs prioritization of constructs for the framework.

In the behavior and relationships assessment stage, participants were asked the following questions:

- Are there any of these factors where, if it's extreme enough, this factor alone would change your decision about using AI, regardless of how the others look? (Clarifying example if participants needed clarification: high consequence of error might set the limit on how much AI can be involved.)
- Do any of these matter more for certain types of tasks than others?
- Do any of these feel like they could go together or influence each other?

💡 How does it inform my project?

The first question would surface priority logic for the framework. The second question would surface task dependency. The third question would surface construct interdependency.

In the completeness assessment stage, participants were asked the following question:

- Is there any other factor you consider when deciding about AI use that is not shown here?
- Are there any other factors that would stop you from using AI at all before you even consider how to use (Clarifying example if participants needed clarification: a client contract explicitly states their data may not be applied to AI tools.)

💡 How does it inform my project?

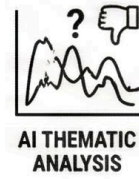
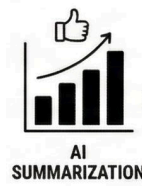
The first question ensured that the construct set is not restricted to what the rapid literature scan identified and allowed discovery of potential constructs that the literature scan might have missed. The second question helped identify factors that gate AI use completely, which would serve as entry logic for the framework.

Task Cognitive Demand



how much active reasoning, interpretation, and mental effort a task require from the UX researcher

Current AI Capability



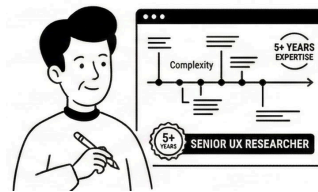
how effectively and accurately current AI tools can actually perform a specific UX research task

Output Verifiability



the ability for the UX researcher to check the AI's outputs and catch errors

Domain Expertise



the level of knowledge the UX researcher has regarding the task at hand

AI Literacy



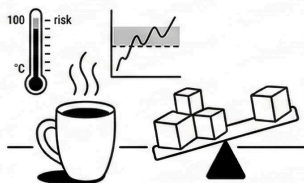
the researcher's skill in using AI tools effectively

Transparency Expectation



how much stakeholders expect the UX researcher to disclose how AI was involved in the research process

Consequence of Error



the amount of damage an error in the research output would cause

Figure 6: candidate construct cards

Construct	Low example	High example
Task Cognitive Demand	Scheduling participant interviews	Synthesizing across multiple interview to identify key user needs
Current AI Capability	Moderating a live interview	Transcribing recorded interviews
Output Verifiability	Checking AI-generated personas (no single correct answer)	Checking an AI-generated summary against the original transcript
Domain Expertise	A UX researcher conducts customer interviews for the first time	A UX researcher who has run many discovery projects at Bynder would be better at spotting flaws in AI-generated interview questions
AI Literacy	Using default prompts and accepts generic output	Knowing how to iteratively prompt AI to refine output quality
Transparency Expectation	Internal brainstorm sessions where nobody asks about methods	Product team wants to know how insights are generated for building an important feature
Consequence of Error	Mistake in a quick internal meeting summary	AI provides flawed insights for a major customer

Figure 7: examples of candidate constructs

3.1.5 Data Analysis Process

Since the research sessions have a two-part structure, the data collected from the inductive exploratory interviews and the deductive construct validation sessions were analyzed separately first before synthesizing them together. This analysis approach was used because data from exploratory interviews and construct validation sessions serve different purposes analytically for the project. Exploratory interviews surfaced potential AI adoption decision factors and other insights from practice and construct validation sessions tested pre-identified AI adoption decision factors against realistic UX research practice. Comparing, contrasting and synthesizing insights from both parts eventually ensures that the end framework and prototyped decision support system are rooted in both practitioner experiences and theories. All data coding was conducted in FigJam. The data analysis process is divided into the following steps:

Step 1: Familiarization Of Raw Data

All research sessions were recorded and transcribed with participant consent. I reviewed the raw transcripts of research sessions first to develop familiarity with the data before diving deep into formal analysis. I took note of recurring points of emphasis during this process without yet applying any formal code.

Step 2: Analysis Of Exploratory Interviews (Data Coding)

The exploratory interview part of the research session was analyzed first before analyzing the construct validation part. Braun and Clarke's (2006) thematic analysis approach was applied. Initial codes were generated across three general areas that the exploratory interview covered as starting points: current AI use, encountered challenges when using AI and AI adoption decision factors that emerged or explicitly mentioned by participants before exposing them to pre-identified, theoretical factors. Specific code labels emerged from the data as the coding process progressed. These initial codes were brought into the synthesis stage (Step 4) for thematic clustering together with the construct validation results. Findings from the coding activity in this step mainly served two purposes. First, they addressed Sub Research Question 1 by making sense of current AI use and challenges. Second, they enable comparison with the pre-identified, theoretical constructs.

Step 3: Analysis Of Construct Validation Sessions (Data Coding)

The construct validation part of the research session was analyzed on the basis of the seven candidate constructs (see Chapter 2) as predefined tags. For each construct, data was coded structurally across three dimensions in correspondence to the structure of the construct validation session, which are relevance (whether the participant confirmed the decision construct is or should be relevant when making AI adoption decisions, and the constructs that are prioritized over others), behavior and relationships (whether the construct is identified as a deal-breaker, whether how it is evaluated depends on the task type, and which other constructs that it goes together with or influences), and completeness (additional factors that are not included in the seven pre-identified constructs, and whether there is any gating factor that stops AI use entirely). The results from the coding activity serve as UX research practitioner validation of AI adoption decision constructs. They directly informed the specific constructs that would go into the AI adoption decision framework as well as how they are integrated into the framework, which addressed Sub Research Question 2.

Step 4: Insights Synthesis

The results from step 2 and 3 were compared and contrasted to synthesize overall insights from the research sessions across two dimensions of comparison. First, the decision factors that emerged from the exploratory interviews were compared against the construct validation results to identify alignment and gaps between reality and theoretical constructs regarding AI adoption decision making. Second, within the construct validation session itself, how participants perceive and apply individual construct in reality is compared against how the

literature defines it. The insights were organized into the three-level organizing structure of the framework (task, human and organizational levels) along with an additional category of themes that did not belong to the existing structure. The confirmed constructs and other design requirements for the decision support framework and the prototyped tool would emerge from these synthesized insights, which address Sub Research Question 3.

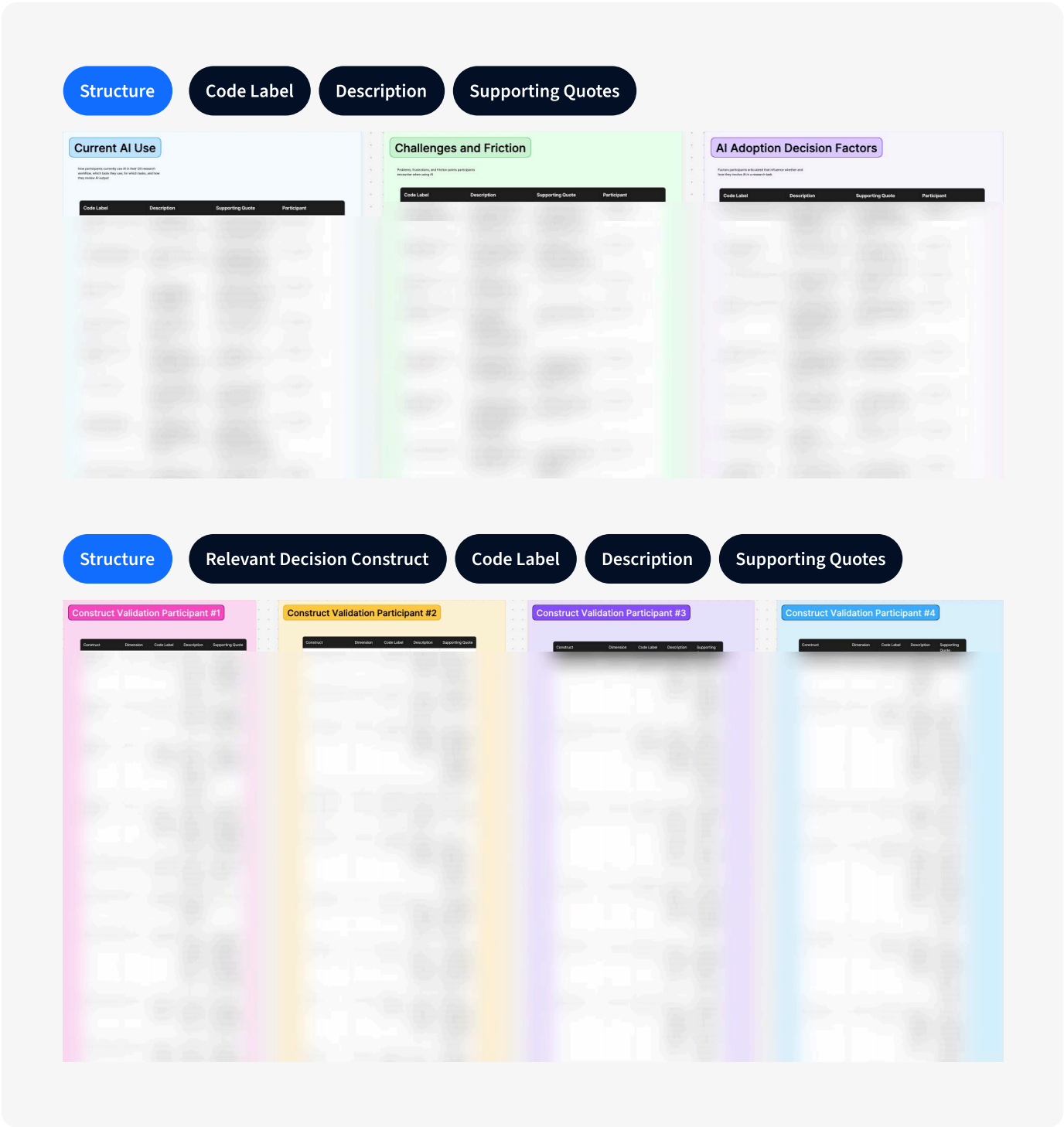


Figure 8: data coding structure on FigJam (step 2 and 3)

3.2 Research Insights

This section covers synthesized insights from the exploratory interviews and construct validation sessions. Insights synthesis follows the process described in step 4 of Section 3.1.5. This section starts by introducing the current state of AI user at Bynder (Section 3.2.1), which sets up the context in which the framework operates. Then Section 3.2.2 presents the construct validation results, in which factors that emerged inductively from exploratory interviews are also compared against the construct validation results to identify alignment and gaps between reality and theoretical constructs regarding AI adoption decision making. Section 3.2.3 presents additional factors and insights that emerged beyond the pre-identified candidate constructs. Lastly, Section 3.2.4 concludes with the synthesized design requirements that would be brought into the framework and prototyped tool design phase.

3.2.1 Current AI use and Challenges in UX research at Bynder

The main goal of the exploratory interviews was to reveal how AI is currently used by UX researchers at Bynder and the challenges that they encountered. They surfaced the following insights that would inform the framework and prototyped tool design.

The exploratory interviews showed that organizational level expectation of using AI exists. All UX researchers actively use AI in their workflow but the motivation, extent and sophistication vary greatly. One participant described using AI across nearly every part of the workflow: "My use of approved AI tools is usually high". In comparison, others prioritize traditional, hands-on analysis. Rather than adopting AI blindly, Bynder's culture encourages strong human-in-the-loop critical thinking on AI use. This reflects a healthy skepticism that serves as a natural defense against AI hallucination. It also ensures that data remains accurate and vetted by humans.

Design Implication

AI use is prominent among UX researchers at Bynder. However, the individual UX researchers' AI adoption levels lie on a wide spectrum, which the framework must accommodate.

All UX researchers stated that they verify AI outputs before using them. Because they work closely with customers, they have a strong feel for the data and can spot errors easily. One participant mentioned "We are really there conducting this thing so it's also easy to spot when the AI is stating something that is not true." Some researchers have even taken a step further to build their own AI output verification mechanisms. One participant mentioned "I ask AI to create a little table of the hypothesis and the objectives I had and categorize the questions next to it. There's always a reference that it isn't lying."

As the team continues to grow, there might be scenarios in which the UX researcher has to verify output from the data they didn't personally collect. The verification approach would also be dependent on firsthand involvement in the interviews, as one participant described that "So far I never was in the position of not being the person interviewing and then having to review the AI output. That would be different. Then I would have to think about another method." To future-proof the UX research quality, there is an opportunity to translate those individual AI output verification mechanisms into a standardized framework that takes into account specific scenarios.

Design Implication

There is a design opportunity for the framework to provide structured, standardized AI output verification approaches for UX researchers. What's more, the UX researcher's distance to the source data (firsthand involvement or not) needs to be built into the framework since it affects AI output verification capability of the UX researcher.

All participants mentioned saving time as a key reason for using AI. One participant explicitly said that speed is the only motivation: *"Why I use AI is to be faster, right? And it's also why we're expected to use AI, I think, not only in this company, but in general."* Another participant mentioned that they use AI in time-consuming tasks: *"For repetitive tasks or analyzing data and all of that, that's where I go straight to the AI because those are the tasks that would take up a lot of time."* The data shows that speed does not change the UX researcher's judgement about appropriate AI use for a task, but rather it serves as a motivation. As a result, this finding shows that although responsible AI integration is the key goal of the tool, it should not add significant time to the UX researcher's workflow.

Design Implication

Speed is a relevant factor when considering adopting AI, which means that the framework should only surface necessary information and avoid redundancy.

3.2.2 Construct Validation Results

The seven candidate constructs were presented to all participants during the construct validation part of the research session. The specific method was outlined in step 4 of the data analysis process (Section 3.1.5). Results are organized by the three levels of the framework (Task, human and organizational level).

Task-level Constructs:

Task Cognitive Demand was confirmed as one of the top three prioritized factors by 75 percent of the participants. One participant ranked it as their single most important AI adoption decision factor.

Before exposing the participants to the candidate constructs, all participants had surfaced task cognitive demand as a AI adoption decision factor. One participant stated that task repetitiveness is the primary driver for using AI: "The more repetitive that the task is, the better for me is to use AI". One participant said that they use AI for scaffolding and large-scale synthesis, while another participant said that they use AI for uninteresting work that drains a lot of their energy. During the construct validation session, One participant also explicitly confirmed that the task cognitive demand construct was "very in line with the examples that I gave you before."

However, there was a notable gap regarding the original framing of the theoretical construct, which treated cognitive demand as a single intensity spectrum from low to high. Participants revealed that there are different types of tasks with high cognitive demand, but they would be using AI and verifying the output differently for those tasks. One participant said that using AI for large-scale data synthesis (structured in nature) works well: "AI is really, really, really, really capable of doing this, to summarize large chunks of data." In comparison, another participant sees certain high cognitive demand tasks that are generative and interpretive in nature as something they don't want AI to replace: "Those things are much more creative."

Design Implication

The framework should operationalize task cognitive demand as task type because participants prioritized task type rather than cognitive intensity when it comes to making AI adoption decisions.

Current AI Capability was confirmed as a relevant factor by all participants, notably that one participant ranked it as the single most important factor: "This is for me the deciding factor — current AI capability."

The theoretical construct defines current AI capability as something to evaluate upfront when deciding whether to use AI for a task. However, the research study showed participants evaluate current AI capability through firsthand, repeated use of specific AI tools, and they cannot assess AI capability as an abstract concept upfront. Instead, they learn it and understand the capability by using the AI tool. For example, one participant mentioned that "I couldn't instruct the company's approved AI tool in such a way that it would give snippets of customers," and another participant stated that they switched back to manual processes when AI did not complete a thorough processing of a document. " This pattern shows that AI capability is learned through firsthand use. As a result, UX researchers who are unfamiliar or lack prior experience using a specific AI tool for a specific task type lack experiential basis for evaluating the output. Appropriate dependance on automation requires trust built through repeated firsthand experience with a specific tool on a specific task (Lee & See,2024). In this situation, they cannot trust the output and thus have to verify the output more carefully.

Design Implication

Current AI Capability is reframed from a per-task upfront AI adoption decision construct in theoretical literature to a risk condition in the empirical framework: when a researcher lacks prior experience with the specific AI tool for the specific task type, AI output verification is intensified.

Output Verifiability was confirmed as one of the top three prioritized factors by all participants. One participant identified it as a deal-breaker: "If I can't really verify if AI is correct, then I can't really use the data that it gave me."

This construct has the strongest alignment across the entire candidate construct set. Before exposing the participants to the candidate constructs, all participants had already independently described reviewing AI output as a non-negotiable action. Some participants put an emphasis on output verifiability because they have experienced AI fabrication. For example, one participant stated "after uploading a report, AI completely made up an entire document which sounds very convincing."

One participant revealed a finding that output verifiability is not just a fixed property of the task itself, it can also be improved through the user's deliberate interaction and choice with AI: "Our research repository tool gives you a quote of when a user is saying a certain thing and so I've been checking that." Also, as stated in Section 3.2.1, UX researchers lack a structured approach when it coming to verifying AI output.

Design Implication

There is a design opportunity that the framework can suggest structured practices that improve AI output verifiability.

Human-level Constructs:

Domain Expertise was confirmed as relevant by 75 percent of the participants.

The theoretical construct defines domain expertise as an standalone factor in technology adoption decision. However, participants consistently perceived it as a modifier of AI output verification capability rather than a reason to use AI or not. One participant described: "Domain expertise and output verifiability ... if you're really expertise in your domain and you know how a great synthesis or great insight looks like and you know the field, you can also verify the output way more easily. So I think these are a bit connected." The participant also confirmed the connection : "domain expertise is relevant because it makes output more verifiable, whatever the AI gives me I can verify against my own knowledge."

Design Implication

The framework should consider using domain expertise to adjust the guidance on AI output verification intensity.

AI Literacy was confirmed as relevant by all participants.

The theoretical construct considers AI literacy as a factor in AI adoption decision. However, participants reframed it as something more foundational to the use of AI rather than being a per-task decision factor. One participant described it as a pervasive factor: "AI literacy kind of goes hand in hand with everything because like if I am not able to use it and leverage AI effectively because I'm not literate in it, it will affect everything else."

Design Implication

AI literacy was reframed by participants as a background factor for calibrating overall AI use and it's not a decision variable when making task-level AI adoption decisions, so it is excluded from the framework.

Organizational-level Constructs:

Transparency Expectation was not confirmed as a per-task decision factor by any participant.

In practice, no participant actively considers transparency expectation as an AI adoption decision factor. For example, one participant stated: "I'm not considering this at all ... whether they expect me to or not makes no difference to me because I will use it because I need to get a task done." The exploratory interviews did not produce any evidence for transparency expectation as a decision factor either.

Design Implication

Transparency expectation should not appear in the assessment of AI adoption decision.

Consequence of Error was confirmed as one of the top three prioritized factors by 75 percent of the participants.

The theoretical construct states that potential damage from errors influences AI automation decisions. This definition aligned with UX researcher's practice: connected it to product impact: "If it's making a change in the interface that affects a lot of users, it's probably the critical one." Although participants did confirm the relevance of this construct, they showed two different perspectives on how consequence of error influenced their AI adoption decision. One participant described it as a factor that sets absolute limit on AI involvement regardless of what the other factors look like when it's extreme enough: "it trumps everything else if consequence of error was going to be insanely high". Meanwhile, another participant considered consequence of error as something that does not influence the amount of AI involved for a task, but rather influencing the intensity of output verification: "It is not a decision factor when it comes to use AI but is a decision factor when you get to review the output, so like on how much you review the output."

Design Implication

The framework should accommodate both patterns observed in the data, where consequence of error can either cap AI involvement or moderate the intensity of output verification.

During the construct validation sessions, participants were asked to rank the top three most important AI adoption decision constructs. Across all participants, the three most prioritized constructs were Task Cognitive Demand (75 percent of all participants ranked it as a top three construct), Output Verifiability (75 percent of all participants ranked it as a top three construct, deal-breaker status from two participants), and Consequence of Error (75 percent of all participants ranked it as a top three construct). These three constructs carry the most weight in the framework's design and correspond to the three targeted literature review areas (Sections 2.2 – 2.4).

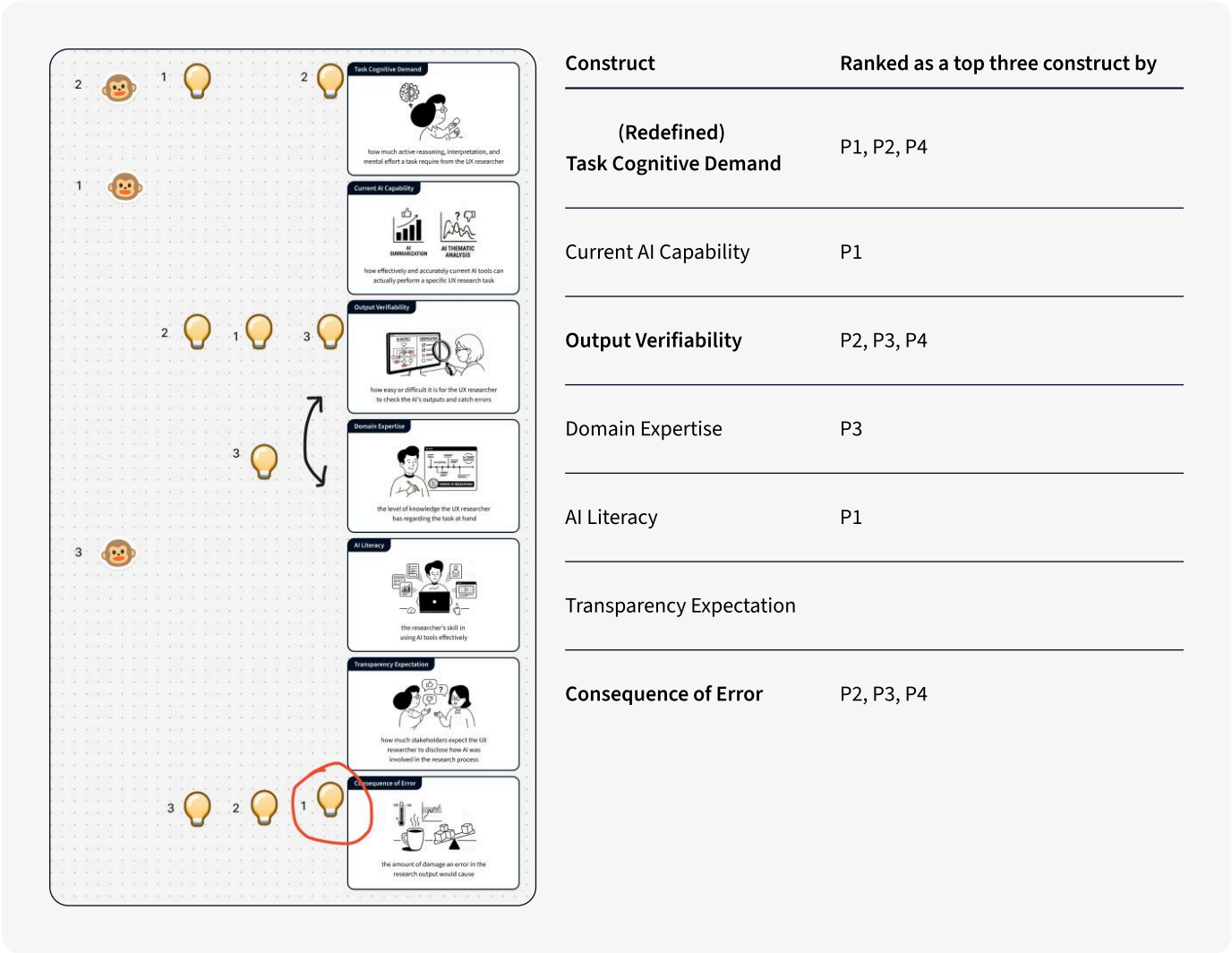


Figure 9: construct prioritization results

The three most prioritized constructs each raised a design tension that the empirical data alone cannot resolve. Therefore, a target literature review for each design tension is needed (Sections 2.2 – 2.4).

For task cognitive demand, the data confirmed its relevance but it revealed a limitation in how the construct was originally defined. The original construct defined task cognitive demand as a single intensity spectrum from low to high, but participants showed that different AI adoption strategies can potentially be applied to tasks with equal cognitive demand depending on the type of cognitive work. Task cognitive demand has a qualitative dimension (cognitive work type) in addition to intensity. As a result, targeted literature review is needed on task typologies that decompose cognitive tasks into distinctive types , so the framework can adjust the AI adoption strategy accordingly (Section 2.2).

For Consequence of Error, the results surfaced two distinct roles of the construct: one participant treats it as a

cap on AI involvement when the consequence of error is high enough for a task, while another participant treats it as a modifier that affects verification intensity without affecting AI involvement itself. These represent two different decision mechanisms. However, the empirical data does not provide a theoretical basis for whether both roles coexist, nor does it surface dimensions regarding how consequence of error should be evaluated. As a result, literature review is needed to justify the dual-role of consequence of error and identify dimensions for evaluating consequence of error (Section 2.3).

For Output Verifiability, the results confirmed it as the top priority construct (75 percent of all participants ranked it as a top three construct and two participants identified it as a deal-breaker). The data also showed that current verification practices among UX researchers are unstructured, which surfaced an opportunity for structuring output verification practices. The research data alone does not provide a theoretical foundation for how verification should be structured as a practice, nor does it identify what verification approaches can be suitably applied to which types of tasks. As a result, literature review is needed to establish a theoretical foundation for structured verification approaches (Section 2.4).

3.2.3 Emergent Constructs Beyond the Original Candidate Construct Set

During the completeness assessment stage of the construct validation sessions, some emergent constructs surfaced beyond the original seven pre-identified constructs.

Data confidentiality and organizational tool access emerged as gating constructs for using AI that fall on the organizational level. Regarding data confidentiality, one participant described: "I cannot of course put customer interviews and so on in a consumer-grade, non-enterprise AI platforms. It has to be tools that Bynder has a contract with or pay for and that don't use customer information to train their LLMs." Regarding organizational tool access, another participant stated: "Legally we can only use the company's approved AI tool, right? So that is something that I consider. I don't go to like other AI tools that maybe would be better." A participant also confirmed this by stating that they build workflows exclusively around tools approved by the information security team.

Design Implication

In the framework, data confidentiality and organizational tool access should be listed as gating condition check before any other assessment since they could block AI use entirely regardless of the other constructs.

Participants also surfaced distance to the source data as a factor that affects verification capacity and thus it's considered as an emergent construct, too (Section 3.2.1). What's more, speed as a relevant factor. However, as discussed in Section 3.2.1, speed serves as a design constraint on the tool rather than a decision construct within the framework and is therefore not included in the final construct set of the framework.

3.2.4 Reframing AI Adoption Decision Process

The research session findings also revealed that the AI adoption decision at Bynder is not a single binary choice but it is made up of two sub-decisions. The first one is an AI permissibility decision (gating condition check), which is about whether AI use is permissible at all. It is determined by organizational tool access and data confidentiality. The second decision is about AI output verification decision. It means when AI use is permissible, how human AI output verification is shaped to fit the specific task. The binary gating decision can be addressed quickly, the calibration decision is more multi-dimensional and this is also where the strategic design opportunity lies. UX researchers at Bynder are rarely in a position where they make decisions about whether to adopt AI. Instead, they do make decisions on how to adopt it, although currently in an inconsistent way. The research sessions showed that there is an opportunity to build a shared structure to support this second decision that is more complex than the first decision.

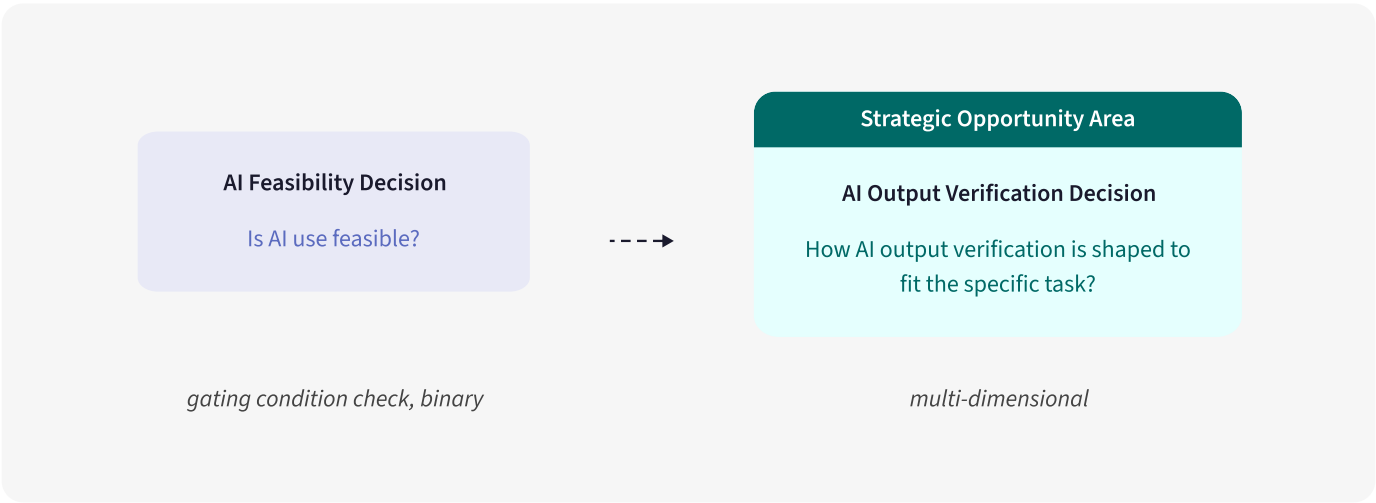


Figure 10: reframed AI adoption decision process

3.2.5 Summary of Empirical Findings

Overall Attitudes on AI and Design Constraints:

Proactive AI adoption : Bynder is actively leaning into the industry-wide technology shift of AI adoption. It is navigating this shift by building an encouraging environment for AI use, which is reshaping UX research practice broadly.

Wide AI adoption spectrum: Individual UX researchers’ engagements with AI lie on a wide spectrum, from those who push AI to its limit to those who engage AI more cautiously.

Speed as a design constraint: Speed is shown to be the universal motivator for AI adoption among UX researchers. Therefore, the proposed tool should only surface necessary information to avoid adding friction to UX researcher’s workflow.

Top Three Prioritized Constructs

- Task Type (Reframed)
- Output Verifiability
- Consequence of Error

Reframed Constructs

- Task Cognitive Demand
- Current AI Capability
- Domain Expertise

Dropped Constructs

- Transparency Expectation
- AI Literacy

Emergent Constructs

- Data Confidentiality
- Organizational Tool Access
- Distance to source data

Strategic Design Opportunity

UX researchers recognize that AI adoption is a natural evolution of their workflows, and responsible adoption relies heavily on output verification rather than restricting access. As a result, there is a clear design opportunity to build a framework that equips researchers with structured, standardized approaches to AI output verification to ensure data integrity.

4 Framework Design

This chapter presents the design of the AI adoption decision framework and the prototype that operationalizes the framework architecture and decision logic. It first consolidates a set of design requirements that inform the design of both the framework and the prototype and then presents a walkthrough of the prototype using an example of a real task in UX researchers' workflow at Bynder.

4.1 Deriving Design Requirements From Design Implications

4.1.1 Consolidation Approach

Design implications are generated in Chapter 2 through targeted literature review and Chapter 3 through empirical research sessions with UX researchers at Bynder. This section consolidates those design implications into a set of design requirements that inform the design of the proposed framework and prototype. Also, framework requirements and prototype requirements are split since the framework addresses the overarching AI adoption decision logic, while the prototype addresses the AI use decision support experience in practice through a functional tool.

4.1.2 Framework Design Requirements

The framework design requirements are grouped into three categories: architecture requirements (how the factors within the framework are organized into a decision structure), content requirements (what is assessed within the framework), and verification requirements (how AI output verification is structured as a component of the framework).

Architecture requirements

Architecture requirements are about the structure in which all the decision constructs in the framework are organized. They are rooted in the three-level organizing structure (Section 2.1.1) and the empirical findings from the research sessions with UX researchers that certain organizational conditions serve as gatekeepers (preconditions) for any AI use (Section 3.2.3).

≡ Framework Design Requirement 1 (F-DR1)

The decision factors in the framework are organized into three levels (task, human, and organizational). Each level contributes differently to the decision flow.

≡ Framework Design Requirement 2 (F-DR2)

The framework incorporates data confidentiality and organizational tool access as the gating layer for AI use before any further assessment.

Content Requirements

Content requirements are about how validated constructs are treated within the framework. They are based on the construct validation results (Section 3.2.2) and the targeted literature review in Section 2.2 and 2.3. They jointly adjusted the original definitions on how these constructs should be accessed.

≡ Framework Design Requirement 3 (F-DR3)

The framework operationalizes task cognitive demand as task type (well-structured vs. ill-structured) to determine which AI output verification approaches are suitable for the given task.

≡ Framework Design Requirement 4 (F-DR4)

The framework addresses consequence of error as a dual-role construct (ability to cap AI involvement or a modifier of verification intensity). It is assessed on severity and reversibility. Thresholds for capping AI involvement are not fixed but left to the researcher's contextual judgement.

Verification Requirements

Verification requirements are about how AI output verification is built into the framework, which is also the core strategic design opportunity area. They are derived from the targeted literature review on verification approaches (Section 2.4) and from the empirical findings that UX researchers lack structured approaches for verifying AI output (Section 3.2.1).

≡: Framework Design Requirement 5 (F-DR5)

The framework provides the UX researchers with structured AI output verification approaches.

≡: Framework Design Requirement 6 (F-DR6)

The framework surfaces risk conditions that modify verification intensity of AI output when these conditions are present.

The framework identifies four risk conditions derived from design implications in Chapter 3: elevated consequences of error, lack of prior experience of using a specific AI tool for a specific task (reframed definition of current AI capability), non-firsthand involvement with the source data, and low domain expertise for the specific task. UX researchers would be signaled to intensify their AI output verification practices, while the UX researcher makes the judgement on the specific verification approach taken.

4.1.3 Prototype Design Requirements

Prototype design requirements define the experience of the prototype that operationalizes the framework. They are rooted in two sources: the DSS design principles identified in Section 2.5 and empirical findings from Section 3.2.1. The decision logic is not restated here since the prototype inherits it from the framework. The requirements below are only experiential properties of the prototype in addition to the decision logic.

≡: Prototype Design Requirement 1 (P-DR1)

The prototype is designed for one-time use per research task.

(It's important to note that a research task is defined as a discrete unit of work within a research project. It does not cover the entire research project. For example, drafting interview scripts and generating insights from a set of user interview sessions are examples of research tasks.)

≡: Prototype Design Requirement 2 (P-DR2)

The prototyped tool makes visible how the outputs are derived from the user input.

≡: Prototype Design Requirement 3 (P-DR3)

The prototyped tool presents its outputs as suggestions rather than prescriptions and encourages the UX's researcher's independent judgement.

4.2 Framework Architecture and Decision Logic

4.2.1 Overview of the Framework Structure

The framework inherits the three-level organizing structure (organizational, task, and human levels) that determines which level each factor belongs to. It then operationalizes these decision factors into a three-layer decision flow (gating, AI output verification calibration, and risk conditions for intensified verification) that determines when each factor and condition is triggered.

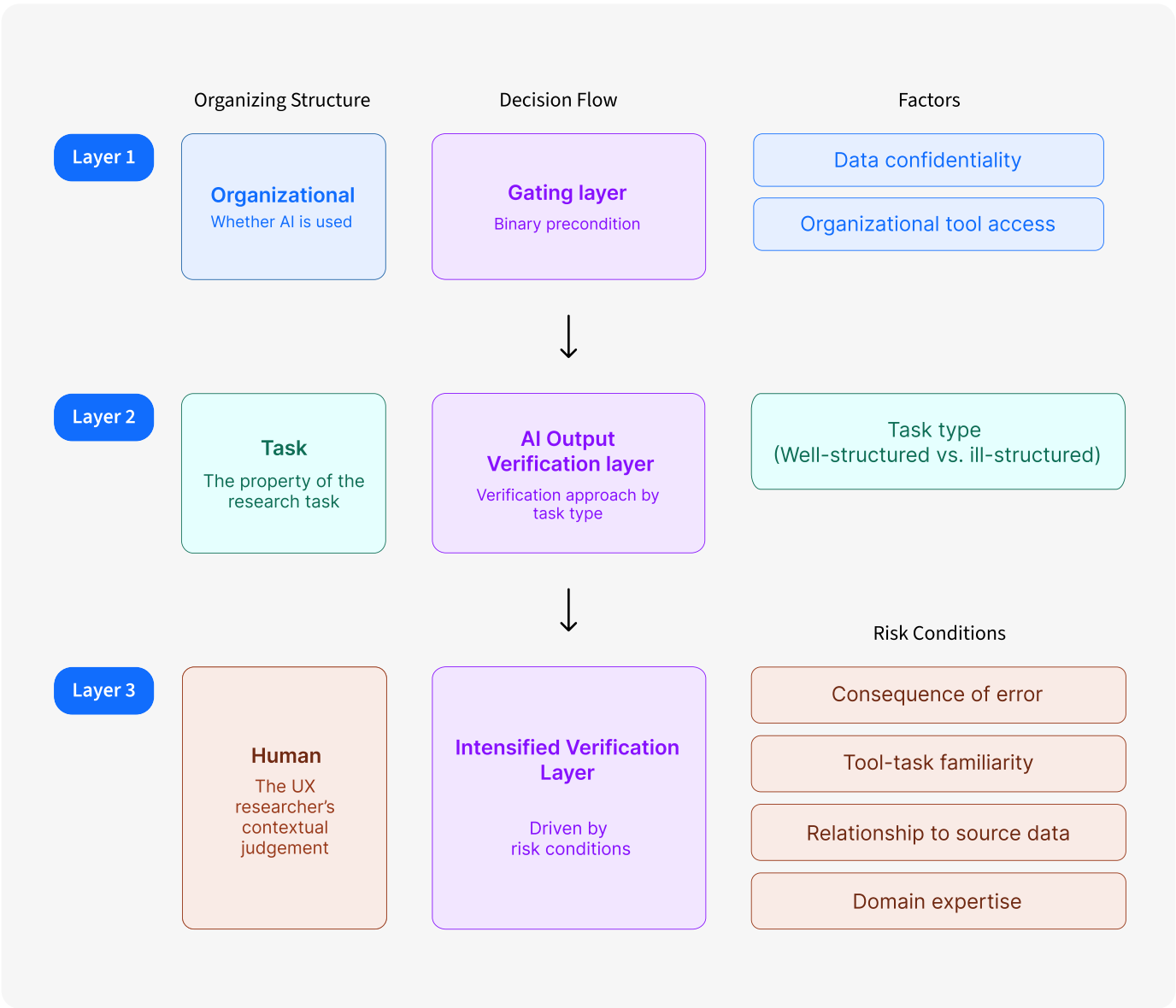


Figure 11: framework structure overview

The gating layer within the decision flow is connected to the organizational level within the organizing structure. The AI output verification layer within the decision flow is connected to the task level within the organizing structure. Lastly, the intensified verification layer within the decision flow is connected to the human level within the organizing structure. The following section will break down the specific components of each layer to showcase the complete decision flow of the framework.

4.2.2 Gating Layer

The gating layer assesses two organizational factors before any other assessment. Data confidentiality and organizational tool access both function as binary preconditions for AI use within Bynder. If either condition fails for the research task at hand, AI use is blocked completely for the UX researcher. If both conditions pass, the task then proceeds to the AI output verification calibration layer.

Data confidentiality determines whether the source data involved in the research task is permitted to be fed into AI tools. Organizational tool access determines whether the use of the available AI tool is approved by Bynder. These two conditions are independent of one another and failure to meet either condition would lead to AI use being blocked for the task at hand.

The gating layer is placed before the rest of the framework because these conditions are treated by UX researchers at Bynder as gatekeepers for AI use in practice. As one participant mentioned it, "Legally we can only use the company's approved AI tools, right?" The framework simply formalizes this gating layer rather than introducing a new logic.

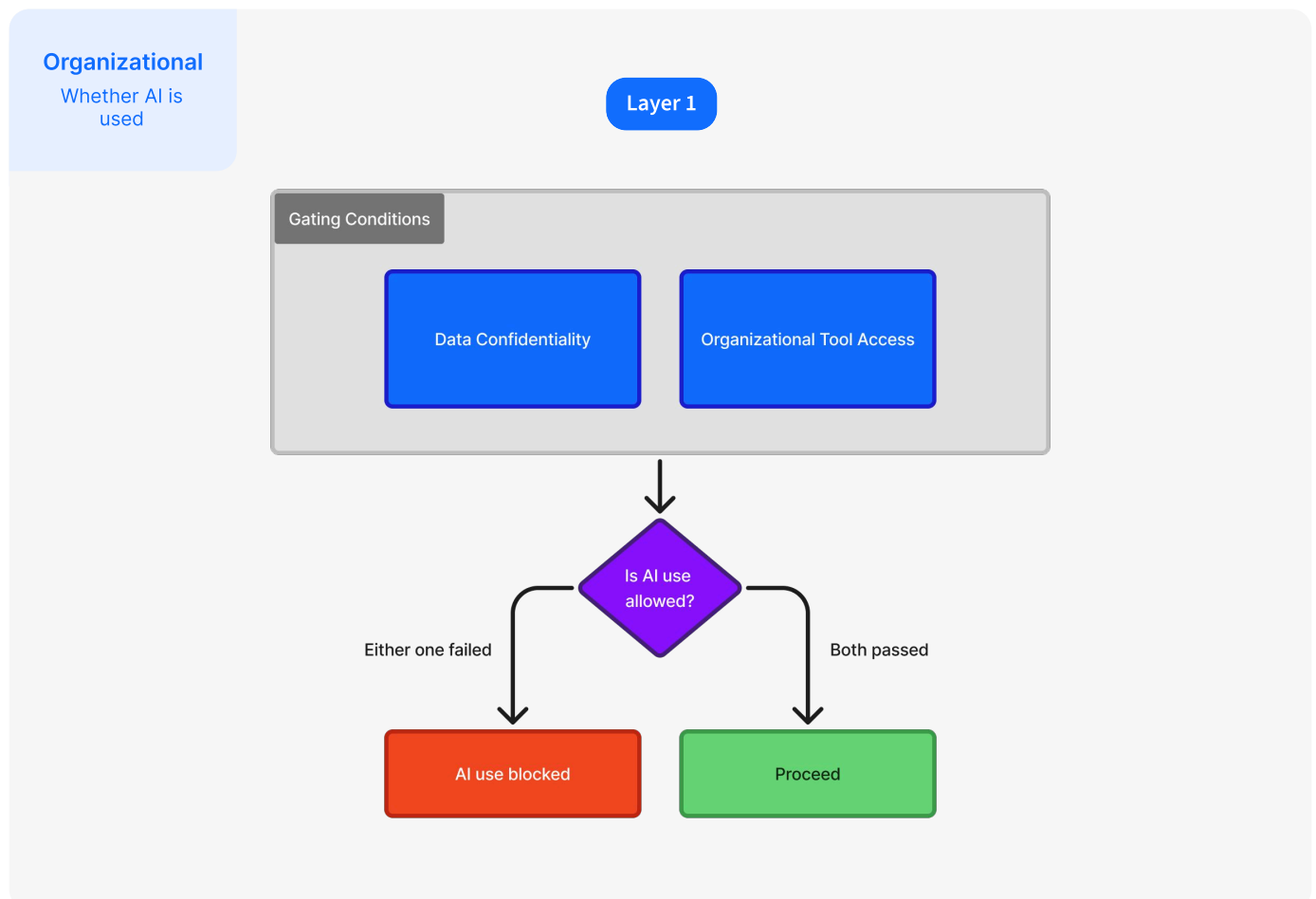


Figure 12: gating layer of the framework

4.2.3 AI Output Verification Layer

The AI Output Verification Calibration layer determines which AI output verification approaches are recommended to the UX researcher for them to apply to a given research task. This layer is connected to the task level within the organizing structure. Task type (well-structured vs. ill-structured) is the input, and the output is the set of AI output verification approaches suitable to the research task. The underlying task type to verification approach logic is rooted in the availability of ground truth as the key condition to effectively detect errors in AI output (Langer et al., 2024) established in Section 2.2.3.

Recap of the Four AI output Verification Approaches

The framework includes four AI output verification approaches (Section 2.2.3):

1. **Source tracing** refers to checking AI-generated output against the original source materials by locating specific data points such as document excerpts and interview quotes to evaluate whether the AI output genuinely represent them.
2. **Criteria-based checking** refers to evaluating AI output against predefined evaluation criteria, where those criteria specify what acceptable outputs should look like.
3. **Comparative verification** refers to the process of checking AI output against an independent knowledge source to surface inconsistencies.
4. **Independent judgement before using AI** requires the researcher to articulate a preliminary interpretation of what the data should show before exposure to AI output, and then use those pre-AI interpretation to compare against the AI output.

Verification Approach vs. Task Type Mapping

The framework maps each verification approach to each task according to two conditions: the approach directly applies to the task type or the approach applies to the task type but only under certain preconditions. The framework does not set any approach as the default starting point. This is an intentional choice since the role of the framework is to support decisions rather than automating decisions (Section 2.3). The framework surfaces the available verification approaches and the UX researcher brings the final judgement on choosing the approach they would like to use.

For well-structured tasks, all four verification approaches apply. Source tracing and criteria-based checking apply directly since they make use of the available ground truth in the source data. Comparative verification and independent judgement before AI use also apply to well-structured tasks. However, they are less suited and efficient as default checks since they verify output through alternatives for ground truth (an independent knowledge source for comparative verification or a pre-AI self interpretation for independent judgement before AI) when the ground truth itself is already available. They can be useful when verification needs to be intensified as they serve as additional checks on the AI output in this scenario.

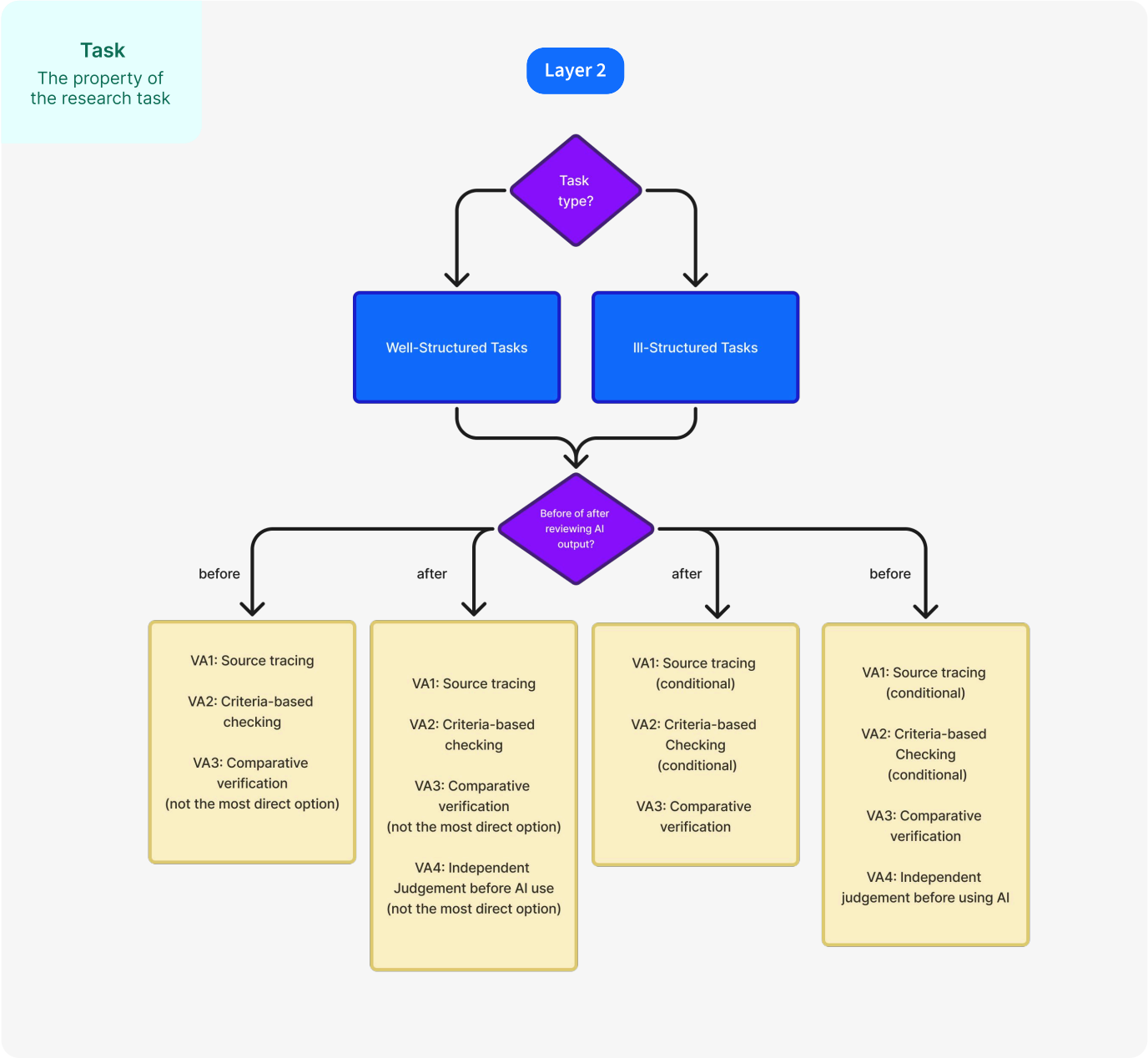
For ill-structured tasks, comparative verification applies directly because there is no external ground truth in ill-structured tasks for checking the AI output against. As a result, comparative verification becomes the most direct verification approach. The other three verification approaches only apply when certain preconditions are met:

Source tracing is only applied to trace the factual elements in the AI output back to the source data, though the interpretive elements in the output cannot be source-traced. For example, an AI generated UX recommendation states that 7 out of 8 participants mentioned (a consensus), therefore (the UX recommendation). In this case, the

factual element that 7 out of 8 participants mentioned (a consensus) can be source-traced.

Criteria-based checking only applies when predefined criteria exist for some evaluable property of the output. While complete correctness criteria is a defining feature for well-structured tasks, other partial criteria, such as coverage and format of the output, can be embedded in ill-structured tasks and thus make parts of the output checkable. For example, reviewing AI generated interview questions against predefined coverage criteria in the research plan (ex. are all topics in the research plan covered?). It only verifies the coverage of the interview questions but there is no complete correctness criteria for the interview questions. Also, it is important to note that not all ill-structured tasks have such criteria and thus this approach is conditional.

Independent judgement before AI use applies only if the researcher has not seen the AI output yet. This precondition exists because this approach prevents anchoring (Bućinca et al., 2021) only if the pre-AI output is self-articulated before exposure to the AI output.



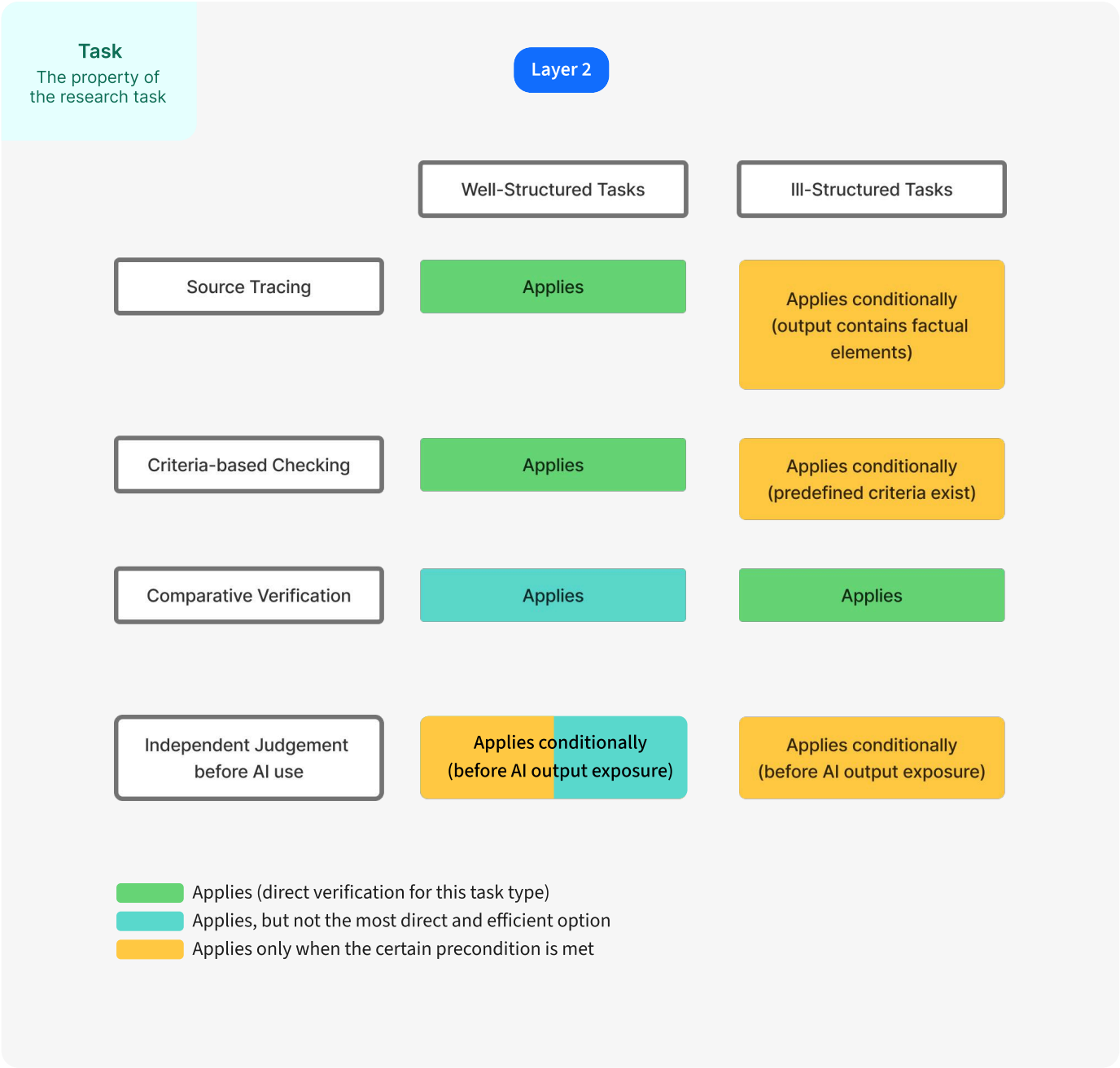


Figure 14: task type to verification approaches mapping

4.2.4 Intensified Verification Layer

The intensified verification layer determines the conditions in which AI output verification needs to be intensified. Four risk conditions are identified in the empirical research insights (Section 3.2). The framework signals the need to intensify and provides a suggestive starting point.

The Four Risk Conditions

The framework includes four risk conditions:

1. **Elevated consequence of error:** This condition is triggered when the task carries high stakes, and it is assessed according to criteria: Severity and reversibility (Section 2.2.2).
2. **Lack of tool-task familiarity:** This condition is triggered when the UX researcher lacks prior experiences or using the specific AI tool for the specific task (Section 3.2.2).
3. **Non-firsthand involvement with the source data:** This condition is triggered when the UX research is not directly involved in collecting the source data (Section 3.2.1).
4. **Low domain expertise:** This condition is triggered when the UX research has limited expertise in the task or project area (Section 3.2.2)

Two of these four risk conditions are placed at a different level than their originating AI adoption decision constructs identified in literature. Consequence of error is placed in the organizational level in the literature scan (Section 2.1.2) because the damage an error would cause depends on the organizational context in which the UX research output is used. Tool-task familiarity, which is a reframing of current AI capability, is placed on the task level because it refers to the realistic performance of current AI tools on given research tasks. In the framework, when these two original constructs are operationalized as risk conditions, they are both evaluated by the UX researcher using their subjective and contextual judgement about the specific research task instead of staying fixed as organizational or task properties. As a result, these two conditions are placed in the human level because they are factors relevant to the UX researcher's personal judgement on AI use.

These risk conditions share one thing in common: they all increase the risk of an undetected error in the AI output slipping through when triggered. When such condition exists, the appropriate way of addressing it is to intensify verification.

How Intensified AI Output Verification is Operationalized: Depth and Breadth

Depth refers to increase the degree of scrutiny with the already-selected AI output verification approach. Depth as an operational dimension is derived from error detection literature, which states that the more severe the consequence of a mistake is, the greater the scrutiny that the user would apply to AI output. Depth has two dimensions that are in line with the two error-detection modes identified by Langer et al. (2024). The first dimension is coverage, which refers to verifying a larger part of the AI output. For example, when it comes to source tracing, it could mean tracing every claim in the AI output back to the source data, not just key ones. The second dimension is judgement rigor, which means applying a stricter standard when assessing AI output. For example, when it comes to comparative verification, it could mean looking into smaller divergences in the outputs. Which dimension is chosen first to intensify verification depends on how the UX researcher applies the verification approaches in practice. For example, when it comes to criteria-based checking, if the UX researcher has already checked all predefined criteria, then depth is achieved mostly through increased judgement rigor than coverage.

Breadth refers to applying additional verification approaches to the AI output than the ones that are already used, which is a practical way of intensifying verification. Because the AI output verification layer (Section 4.2.3) recommends more than one verification approach for each task type, the UX researcher can choose to apply additional verification approaches to a given task.

The Dual Role of Consequence of Error

Consequence of error is the only one of the four risk conditions that not only intensifies verification, but also caps AI involvement level. The dual role is drawn from the empirical research (Section 3.2.2) but also justified by literature (Section 2.2.2). In the framework, consequence of error is evaluated on its severity and reversibility and the threshold at which capping happens is left to the UX researcher's personal judgement rather than being defined by the framework.

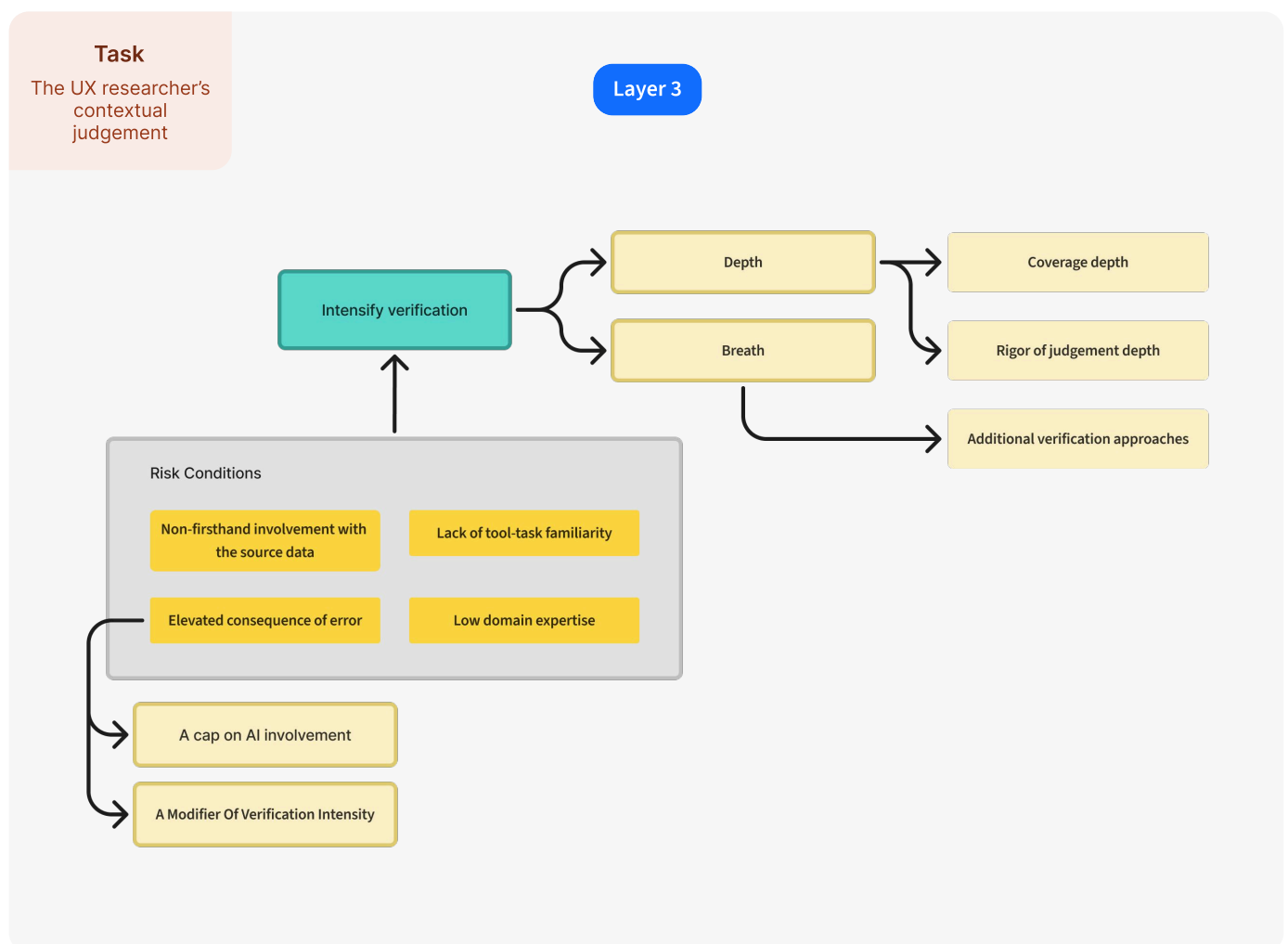


Figure 15: intensified verification layer of the framework

4.3 The Prototype in Operation

This section presents the prototype that operationalizes the framework architecture and decision logic described in Section 4.2 into a tangible tool that UX researchers can use in their workflow. The prototype is not a separate contribution in itself since the embedded decision logic is inherited completely from the framework. The prototype adds the experiential layer of the framework using the prototype design requirements defined in Section 4.1.3: designed for one-time use per research task, making how outputs are derived from user input visible, and presenting suggestive outputs rather than prescriptive.

4.3.1 The Assessment Flow Walkthrough

The prototype presents the assessment flow in five linear steps to collect user input: data confidentiality assessment, organizational tool access assessment, task type assessment, AI output status assessment and risk condition assessment. The five steps are linked directly to the framework's three layers. The first two gating questions correspond to the gating layer (Section 4.2.2), the task type and AI output status questions corresponds to the AI output verification layer (Section 4.2.3), and the risk condition question corresponds to the intensified verification layer (Section 4.2.4).

In order to showcase how the prototype works in practice, the walkthrough uses an example of a real task in UX researchers' workflow at Bynder. The task "Using a Bynder approved AI tool to work across customer interview transcripts to draft UX design recommendations that the team can take action on" is chosen because it is a common AI integration scenario in the UX research workflow at Bynder that happens at the deliver phase of the UX research process. The customer interview transcripts are the source data the AI works with, and the UX recommendations are the outputs that the UX researcher needs to verify.

Step 1: Data Confidentiality Assessment

The data confidentiality assessment belongs to the gating layer and asks whether the data (ex. sensitive customer data) for the task is allowed to be shared with external AI tool, which is a binary choice between permitted and restricted. If it is restricted, AI use is blocked entirely. This assessment sits in the gating layer.

In the case of the example, it assumes that the interview transcripts does not contain any confidential information and thus data confidentiality assessment passes.

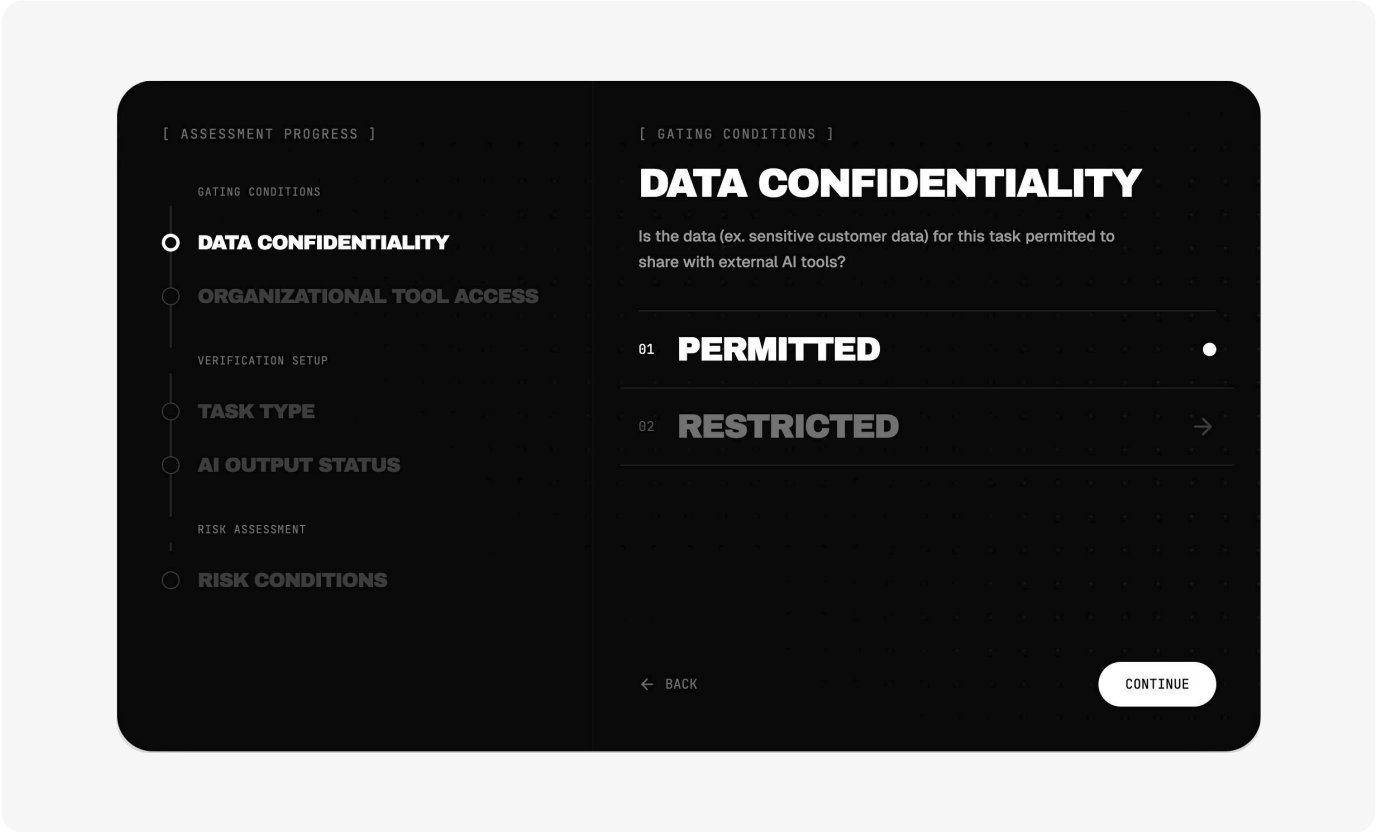


Figure 16: data confidentiality assessment (user interface)

Step 2: Organizational Tool Access Assessment

The organizational tool access assessment belongs to the gating layer and asks whether the specific AI tool the UX researcher intends to use is approved by Bynder for use with work data, which is a binary choice between approved and not approved. If it is not approved, AI use is blocked entirely.

In the case of the example, it assumes that the AI tool used is in compliance with Bynder’s AI tool policy and thus organizational tool access assessment passes.

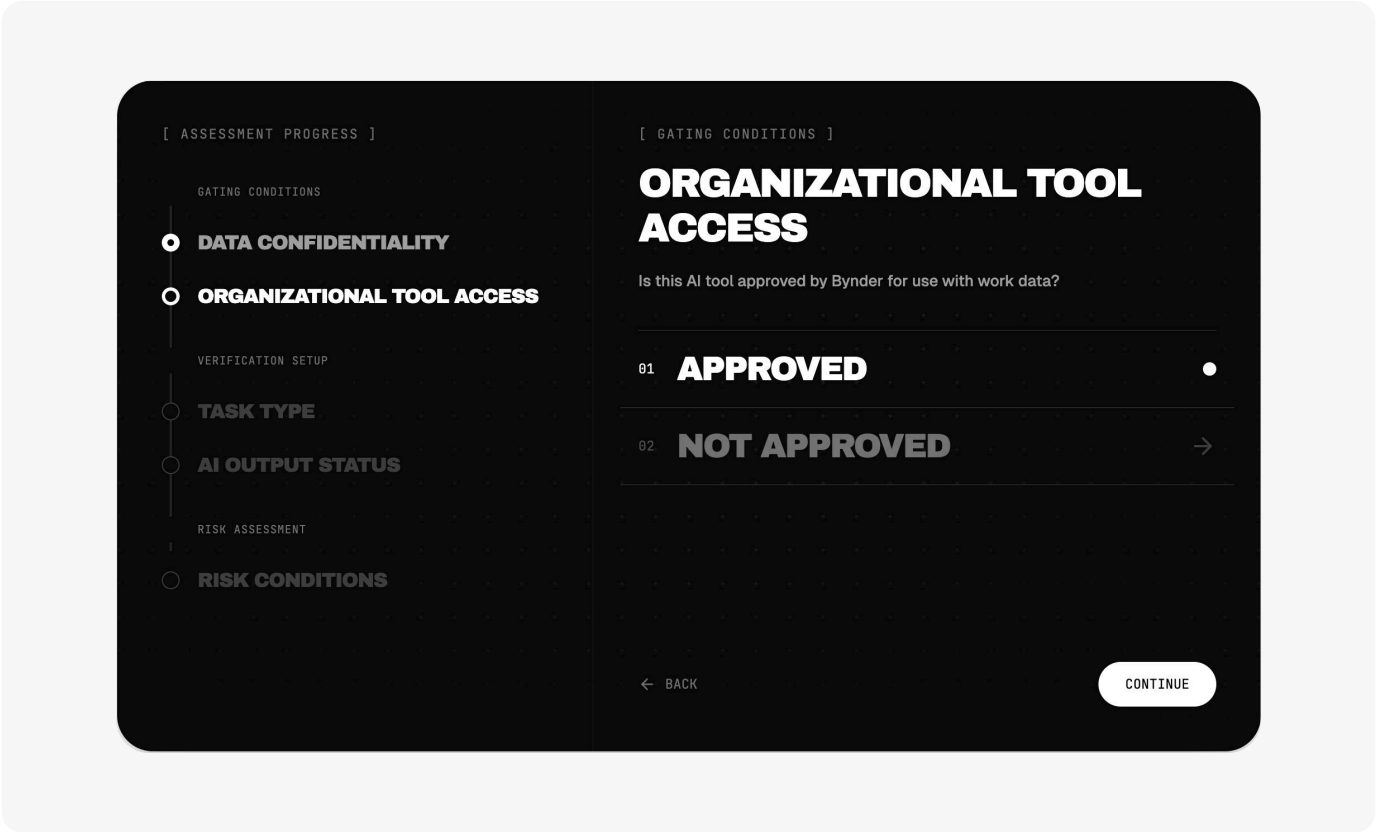
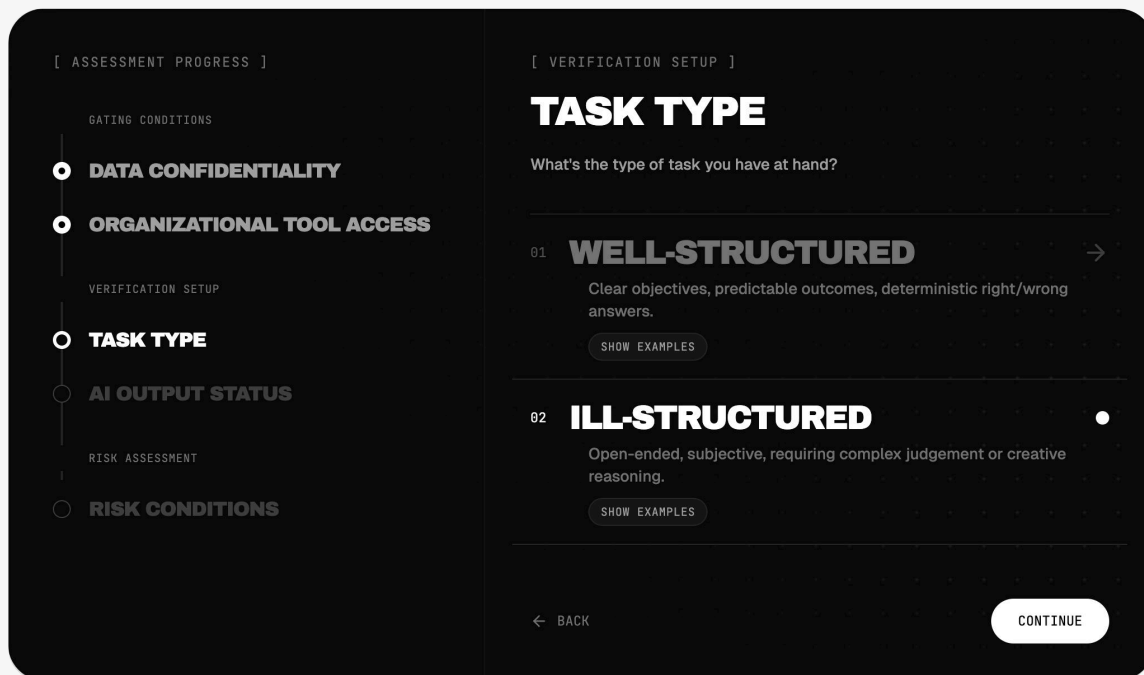


Figure 17: organizational tool access assessment (user interface)

Step 3: Task Type Assessment

The task type assessment belongs to the AI output verification layer and assesses the task type the UX researcher has at hand. It asks the UX researcher to identify the task as well-structured tasks, which are characterized by clear objectives, predictable outcomes and deterministic right or wrong answers, or ill-structured tasks, which are characterized by open-endedness and the requirements of complex and interpretive judgement. This screen includes a feature for the UX researcher to view examples of each task type in practice for them to better understand the task categorization.

In the case of the example, drafting UX design recommendations is an ill-structured task that requires interpretation and thus “ill-structured” is chosen as the task type. It is ill-structured because the parameters around UX design recommendations, such as how to weigh different user inputs in the interviews and what the product team can take action on realistically at a given point in Bynder’s organizational context, are not fixed. Therefore, the resulting UX recommendations are not deterministic, and AI output cannot be simply checked as right or wrong when there is no external ground truth. The task requires strategic judgement and reasoning.



The image shows a user interface for a task type assessment. It is divided into two main sections: a left sidebar and a main content area. The sidebar, titled '[ASSESSMENT PROGRESS]', contains a vertical list of steps: 'GATING CONDITIONS', 'DATA CONFIDENTIALITY', 'ORGANIZATIONAL TOOL ACCESS', 'VERIFICATION SETUP', 'TASK TYPE' (which is highlighted with a white circle), 'AI OUTPUT STATUS', 'RISK ASSESSMENT', and 'RISK CONDITIONS'. The main content area, titled '[VERIFICATION SETUP]', has a heading 'TASK TYPE' and a question 'What's the type of task you have at hand?'. Below this, there are two options: '01 WELL-STRUCTURED' with a right arrow and '02 ILL-STRUCTURED' with a white dot. Each option has a description and a 'SHOW EXAMPLES' button. The 'WELL-STRUCTURED' description is 'Clear objectives, predictable outcomes, deterministic right/wrong answers.' and the 'ILL-STRUCTURED' description is 'Open-ended, subjective, requiring complex judgement or creative reasoning.' At the bottom of the main content area, there are '← BACK' and 'CONTINUE' buttons.

Figure 18: task type assessment (user interface)

Step 4: AI Output Status Assessment

The AI output status assessment belongs to the AI output verification layer and asks whether the UX researcher has already seen the AI output or not for the given task. This answer to this question determines whether the approach “independent judgement before AI use” is shown as a recommended approach because this verification approach is only available if the UX researcher has not been exposed to the AI output yet. Once they have read the AI output, they are not able to articulate genuinely independent interpretation to the task without being biased.

In the case of the example, it assumes that the UX researcher has not seen the AI output yet, and they are using the tool to plan on how they want to be verifying the AI output before using AI.

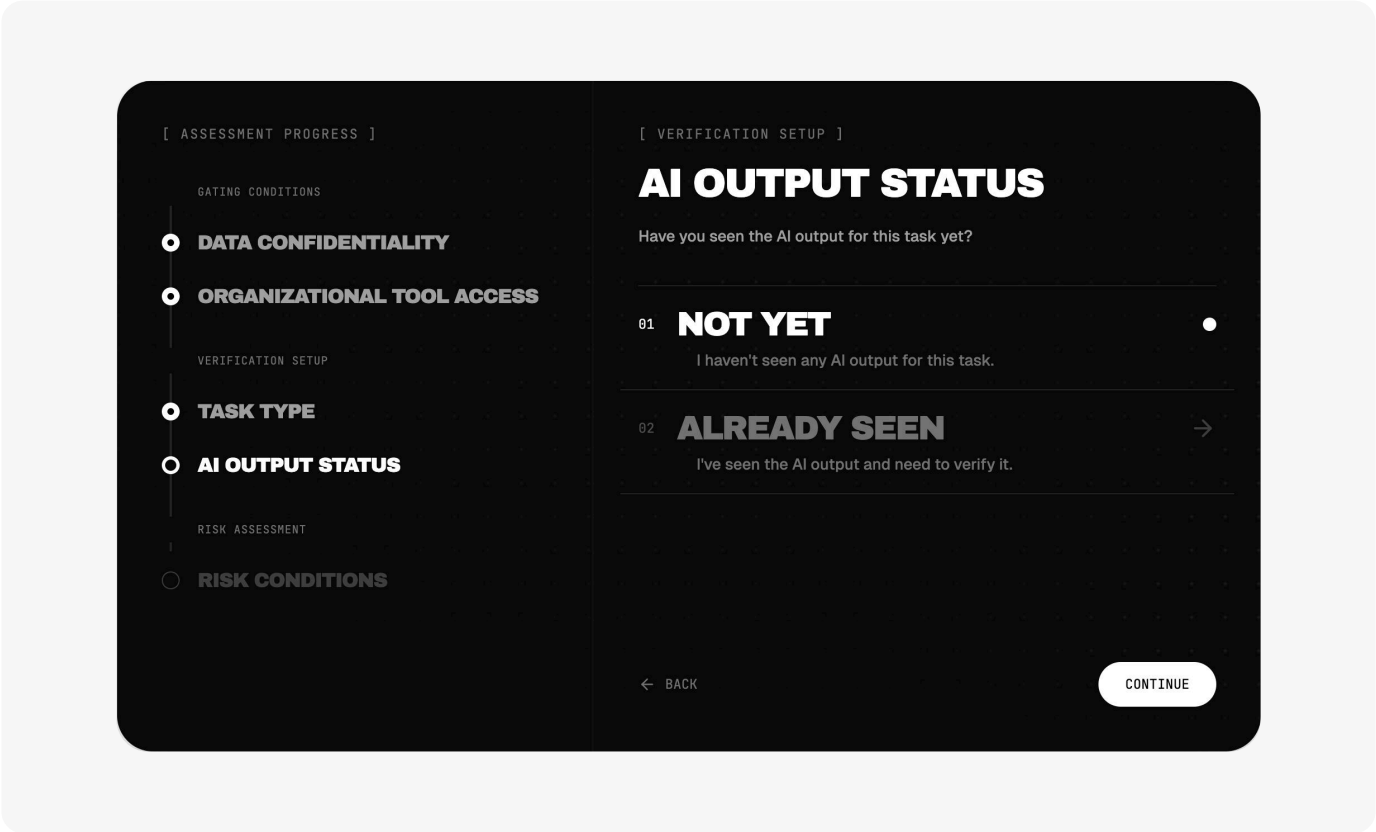


Figure 19: AI output status assessment (user interface)

Step 5: Risk Condition Assessment

The risk condition assessment belongs to the intensified verification layer and presents the four risk conditions (Section 4.2.4) as multi-select options to the UX researcher for them for judgement if any one applies to their task at hand. These risk conditions are positioned as signals for the UX researcher to consider intensifying AI output verification.

In the case of the example, consequence of error is flagged because UX design recommendations could impact product decisions, and it assumes that the severity is high and reversibility is low (two evaluation dimensions of consequence of error).

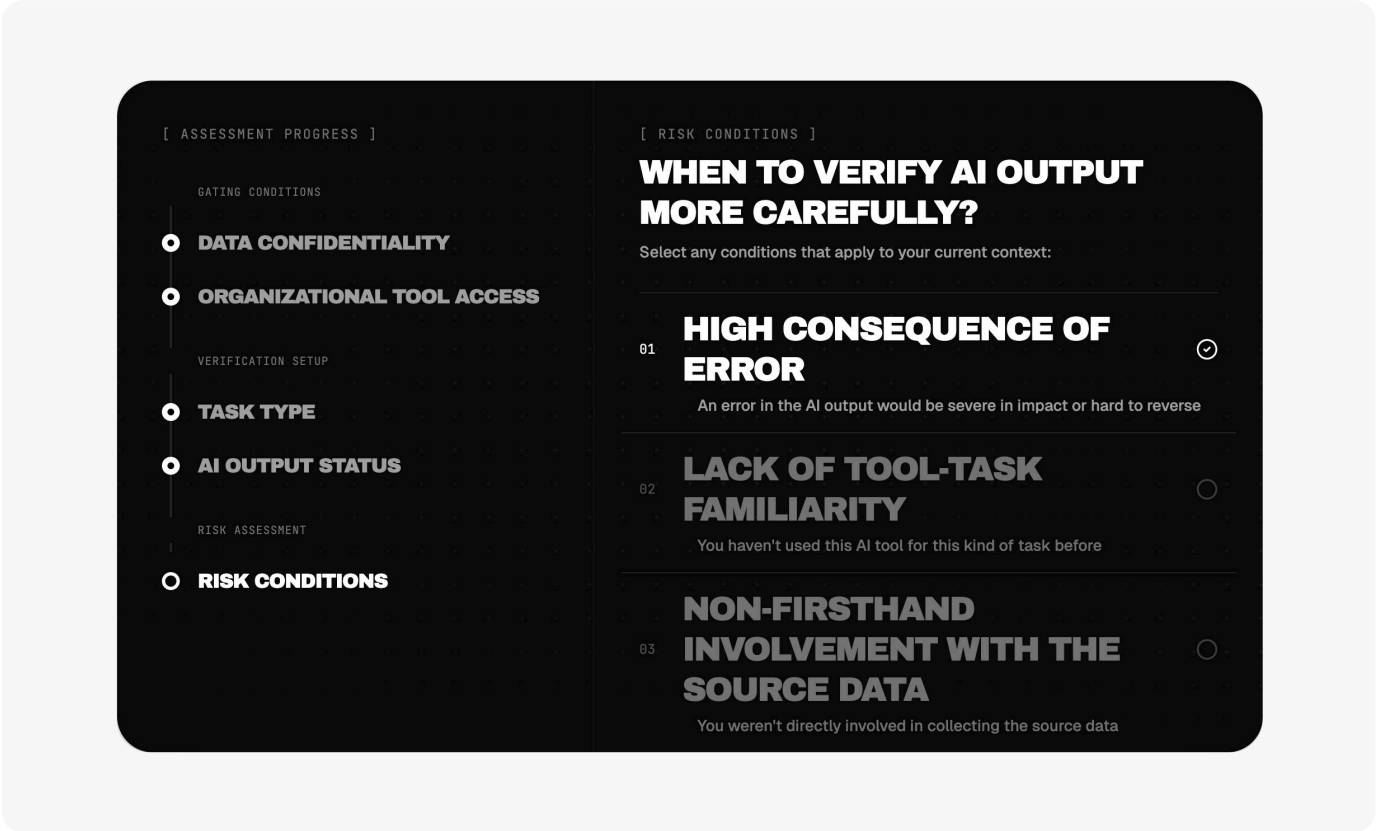


Figure 20: risk condition assessment (user interface)

Result Screen

Once the user has provided their answers to all five assessment questions, they can then proceed to view the result. The tool presents the flagged risk condition and recommended AI output verification approaches based on their answers to the assessment questions.

In the case of the example, because high consequence of error is flagged as a risk condition. The result screen surfaces the dual role of consequence of error established in the framework (Section 4.2), which means that the UX researcher can either intensify AI output verification (consequence of error as a modifier of verification intensity) or reconsider the level of AI involvement (consequence of error as a cap on AI involvement). The tool intentionally only shows that both options exist and does not define the threshold of capping (Section 2.2.2). It leaves the final judgement to the UX researcher, which aligns with the framework's decision-support role and Prototype Design Requirement 3: showing outputs as suggestions rather than prescriptions (Section 4.1.3).

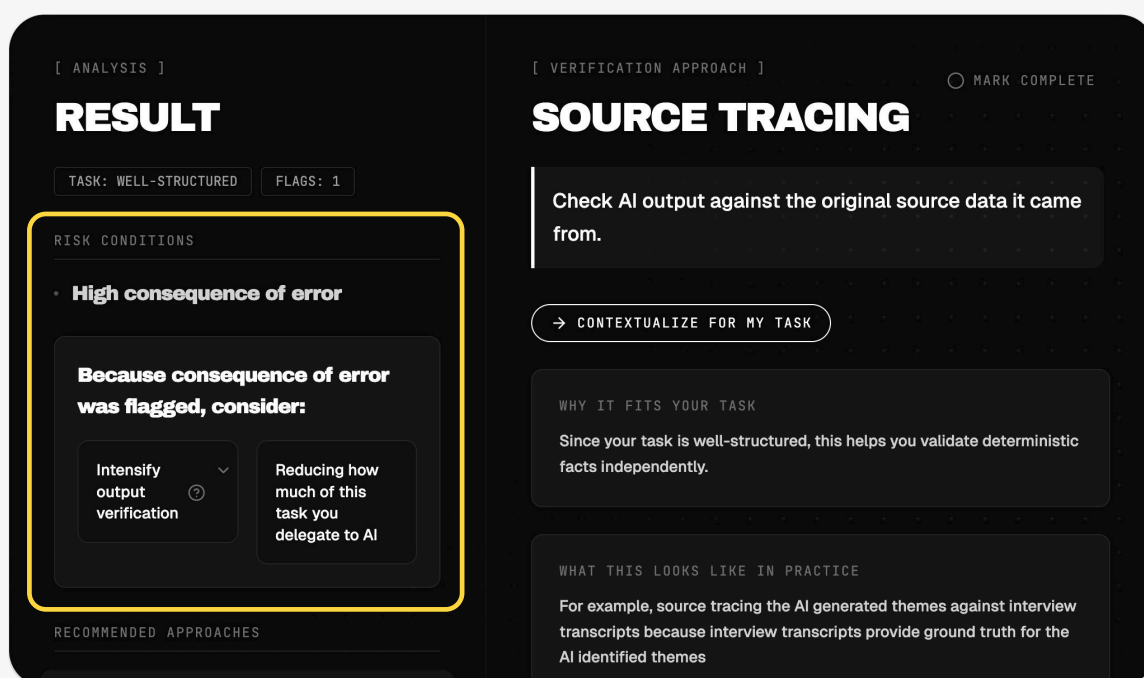


Figure 21: result page (user interface)

Since intensified AI output verification is operationalized by two dimensions: depth (intensify within already chosen verification approaches) and breadth (intensify by using additional verification approaches), the tool also surfaces the two dimensions of intensified verification. The user can click on the info icon next to “Intensify output verification” to open a modal to view the explanation of the two dimensions.

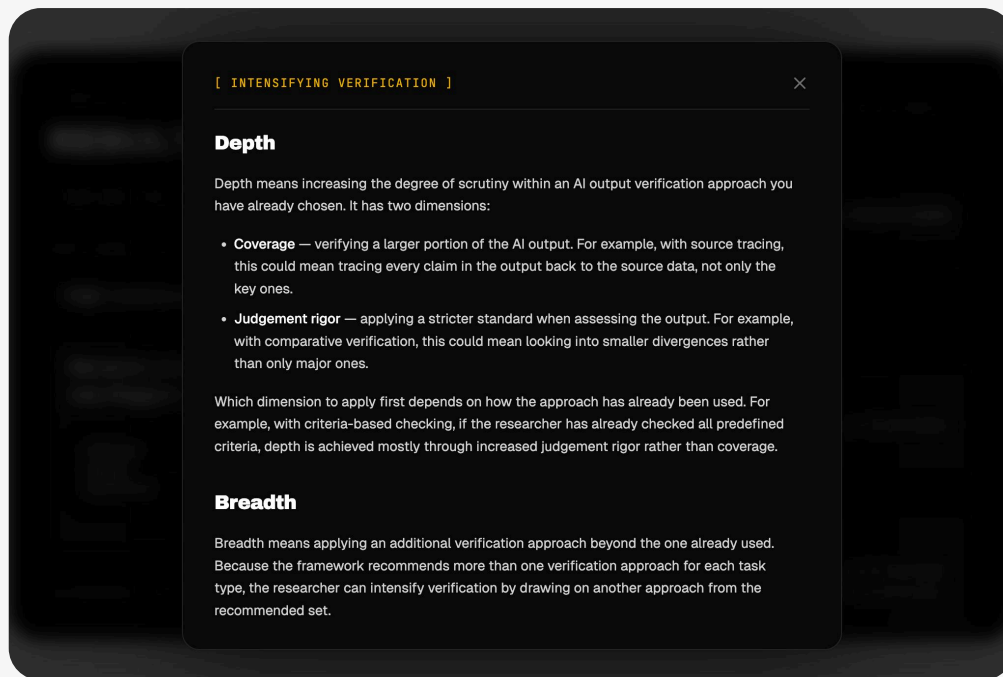


Figure 22: two dimensions of intensified verification (user interface)

In the case of the example, the result screen also presents four recommended AI output verification approaches. The resulting recommended AI output verification approaches follow the reasoning described in Section 4.2.3. Ill- structured task was chosen (Step 3) and the UX researcher has not yet seen the final AI output (Step 4). Therefore, comparative verification and independent judgement before using AI are justified as suitable output verification approaches. Source tracing may be applied but only under certain conditions. For example, when there is a factual element in a UX recommendation that a certain number of participants mentioned a specific issue, this factual element can be source traced to the original transcripts but it does not verify the interpretive UX recommendation itself. Criteria-based checking may also be applied conditionally. For example, when it is specified in the project brief that the UX recommendations need to address all research questions, the coverage criteria can be used to check the output but it does not verify the interpretive UX recommendation itself since there is no complete correctness criteria. The user can hover over the tooltip icon next to the verification approach to view the condition (see Figure 23).

For each verification approach, the user can also view the reasoning for why this approach is recommended, which aligns with Prototype Design Requirement 2: The prototyped tool makes visible how the outputs are derived from the user input. An example of what this verification approach looks like in practice and the materials needed to start with this verification approach are also presented to the user to ensure better understanding (see Figure 24).

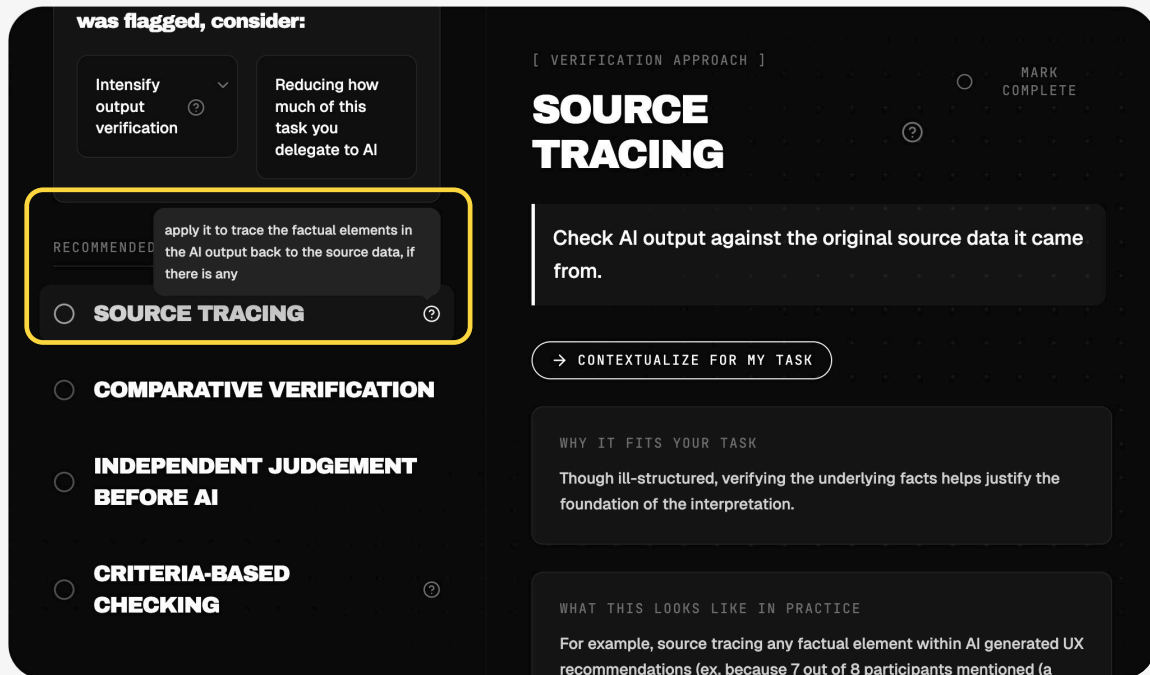


Figure 23: view the condition in which source tracing applies (user interface)

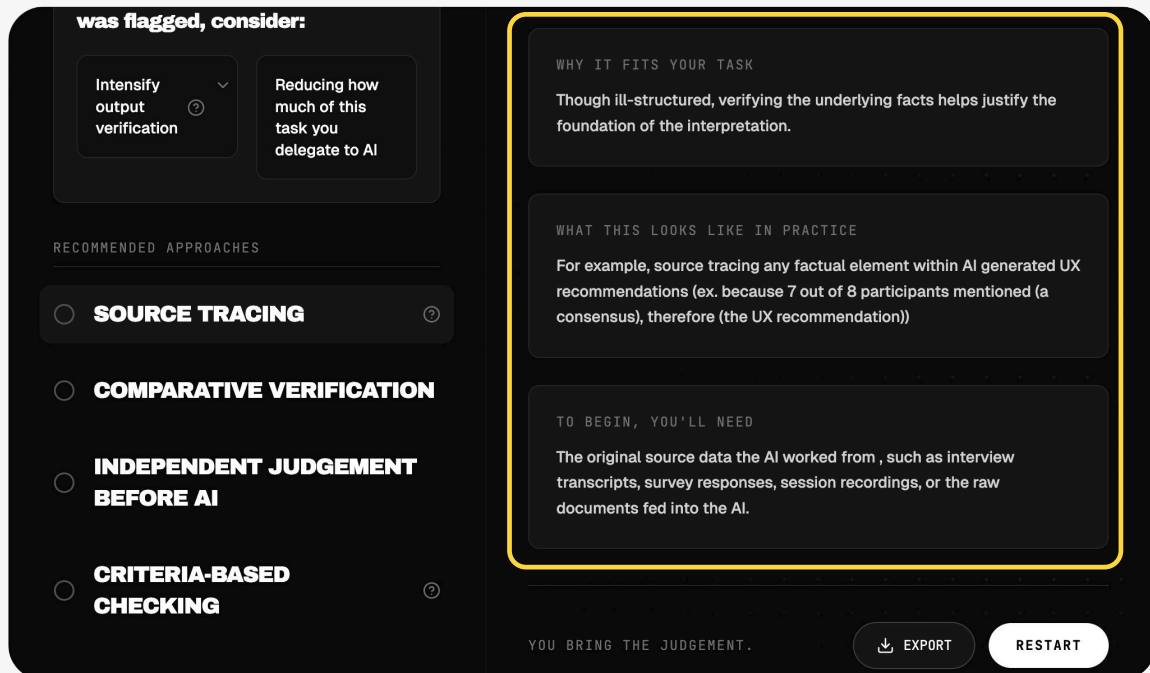


Figure 24: explanatory information of the chosen verification approach (user interface)

4.3.2 The Contextualization Feature

The assessment flow walkthrough demonstrates the core logic behind the prototype and how the framework is operationalized in practice. Additionally, the prototype also includes a contextualization feature that tailors each recommended AI output verification approach to the specific research task the UX researcher has at hand. This feature exists because the verification approaches could still seem abstract to the UX researcher even when example use scenarios are presented to them on the result screen. The contextualization feature opens up a short form, asks for a description of what the UX researcher is using AI for and the data AI works with, and then generates a set of concrete guidance on how to apply the specific verification approach to the specific task using a Gen AI model(gemini-2.5-flash). The guidance includes practical steps, a risk callout, and a boundary statement of what this output verification approach covers and leaves out.

The contextualization feature is an optional feature because it could feel redundant for those who are able to contextualize their tasks at hand directly without external assistance, and the user would only use the contextualization feature as additional assistance only when needed.

The specific system prompt for the Gen AI model applied to each verification approach can be found in the appendix (Page 88). Each system prompt includes definition of the Gen AI role, definition of the verification approach, output structure, content rules, voice rules, edge cases and an example output.

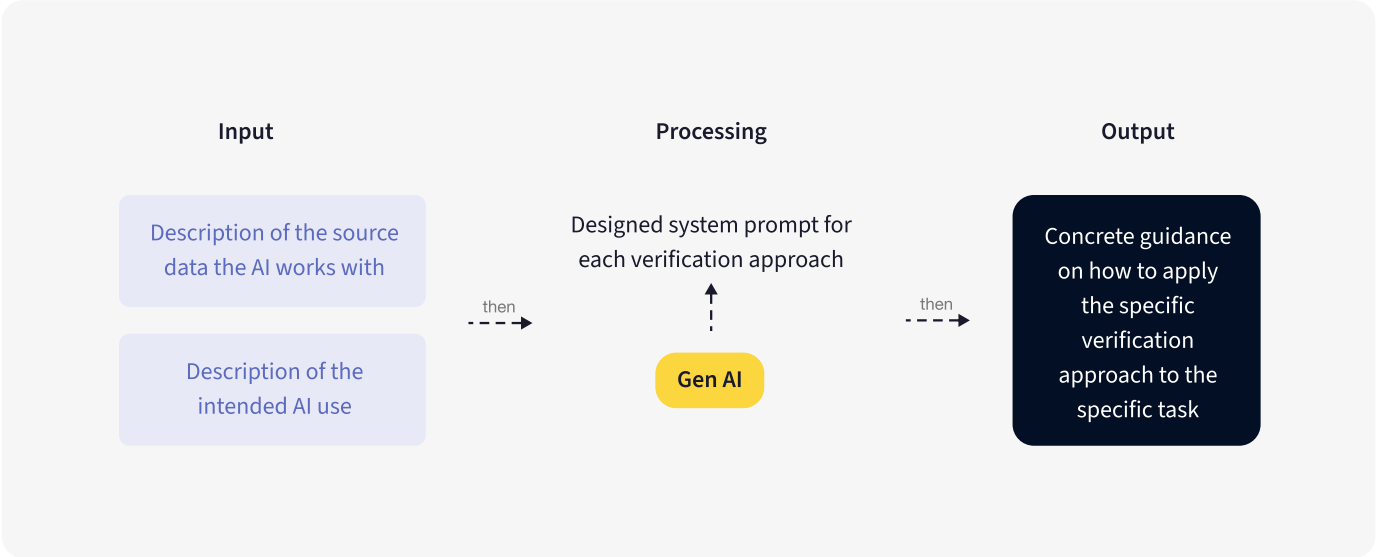


Figure 25: the contextualization feature overview

[TAILOR TO YOUR TASK]

TELL US ABOUT YOUR TASK

We'll use this to make each recommended approach concrete for what you're working on.

WHAT ARE YOU USING AI FOR?

Using a Bynder approved AI tool to work across customer interview transcripts to draft UX design recommendations that the team can take action on 145 / 300

WHAT DATA IS THE AI WORKING WITH?

6 customer interview transcripts

CANCEL GENERATE

Figure 26: the contextualization feature - data input (user interface)

[FOR YOUR TASK: SOURCE TRACING] ×

⚠️ AI GENERATED CONTENT: These are AI generated content, use them as suggestions. Feel free to modify, override, or disregard certain parts using your own judgement.

This Source Tracing approach helps you verify the AI's factual claims by linking them directly to your interview transcripts. Given the high consequence of error in design recommendations, ensuring the AI's foundational understanding is critical.

1. For each AI-generated UX design recommendation, identify the specific customer quotes or transcript sections the AI cites as evidence.
2. Locate these exact quotes within your original 6 customer interview transcripts.
3. Scan the surrounding context in the transcripts to confirm the quote's original meaning and to find additional supporting or contradictory statements related to the recommendation.
4. If the AI doesn't cite specific evidence, manually search the transcripts for keywords or themes related to the recommendation to find supporting data.

Risk Callout: Be vigilant for AI interpretations that misrepresent the original intent of the customer.

Boundary: Source tracing validates the factual basis of the AI's claims but does not verify the AI's interpretive claims, framing, or prioritization of recommendations.

Practical steps

Risk call-out

Boundary statement

Figure 27: the contextualization feature - output (user interface)

5 Evaluation

This chapter presents the evaluation process of the prototype that operationalizes the framework against three areas: clarity, usability, and conceptual completeness. A think-out-loud walkthrough format with two standardized scenarios was used for the evaluation. The chapter concludes by presenting the evaluation results matched to the three evaluation areas as well as the prototype iterations in reaction to gathered insights.

5.1 Evaluation Goals

The evaluation phase assesses the prototyped tool that operationalizes the framework. Because the framework logic is embedded in the prototyped tool, the two are evaluated together. Three areas constitutes the evaluation:

1. **Clarity** - Users should be able to follow each step of the assessment flow in the prototyped tool and understand what is being assessed. They should be able to understand the connection between their inputs and the prototype's final output. The embedded framework logic and how information is organized are clear to the user.
2. **Usability** - The prototype should only surface necessary information and no information should feels redundant.
3. **Conceptual Completeness** - The user should recognize that the AI adoption factors surfaced in the assessment flow correspond to what they would consider in practice and identify any missing factor or irrelevant factor.

5.2 Evaluation Approach

The same participants who have previously joined the research sessions (Chapter 3) participated in the evaluation sessions. Therefore, the same profile documented in Section 3.1.2 is followed: convenience sampling of a representative sample of UX researchers from Bynder's UX research team, 3 to 6 years of experiences in UX research, all had prior experience using AI tools in their workflows though the extent and sophistication vary. This was a deliberate and practical choice given the small UX research team size at Bynder.

Each evaluation session lasted about 30 minutes and used an open walkthrough format in which participants were encouraged to think out loud as they were going through the prototyped tool. In the first pilot session, the participant was asked to apply a recent task in which they used AI to the prototype. This format was initially chosen to surface if and how the prototype would apply to realistic UX research tasks. After that, the following sessions were conducted using two standardized scenarios to contrast the two distinctive paths within the framework (well-structured task or ill-structured task) and ensure better comparability across participants because the task type becomes the main variable.

A short briefing about the current project state and the prototype's purpose was presented to the participant at the beginning of the evaluation session. Then they were assigned one of the two scenarios (well-structured task or ill-structured task). Each scenario also includes an underlying risk condition to test whether it could be correctly identified by the participant. The followings are the two scenarios:

Scenario 1 - Well-structured task

"You're putting together the research report and need supporting customer quotes to include in it. You're using a Bynder approved AI tool to go through 6 customer interview transcripts and find quotes where customers talk about the issues you're reporting on, so you can pull the strongest ones into the report. These interviews were run by a colleague on the team and you're working from their transcripts. You are allowed to feed anonymized customer data to the AI tool." (The underlying risk condition: Non-firsthand involvement with the source data)

Scenario 2 - Ill-structured task

"The team is working on improving the experience of a customer-facing dashboard within the product. You've been asked to come up with UX recommendations based on what customers are reporting in user testing sessions. You're using a Bynder approved AI tool to work across 6 customer interview transcripts, identify what's not working for customers, and draft UX recommendations the team can act on. Your recommendations will feed into the team's design decisions and inform product roadmap. You are allowed to feed anonymized customer data into the AI tool." (The underlying risk condition: High consequence of error)

At the end of each session, the participant was also asked the following follow-up questions to ensure additional insights relevant to the evaluation goal were captured:

Q1: When you were going through the assessment flow, was there anything you couldn't follow or you think that feels confusing? (Further addressing clarity and usability)

Q2: Did the tool surface what you'd have considered yourself, miss something that matters, or flag something you'd treat as irrelevant? (Further addressing conceptual completeness)

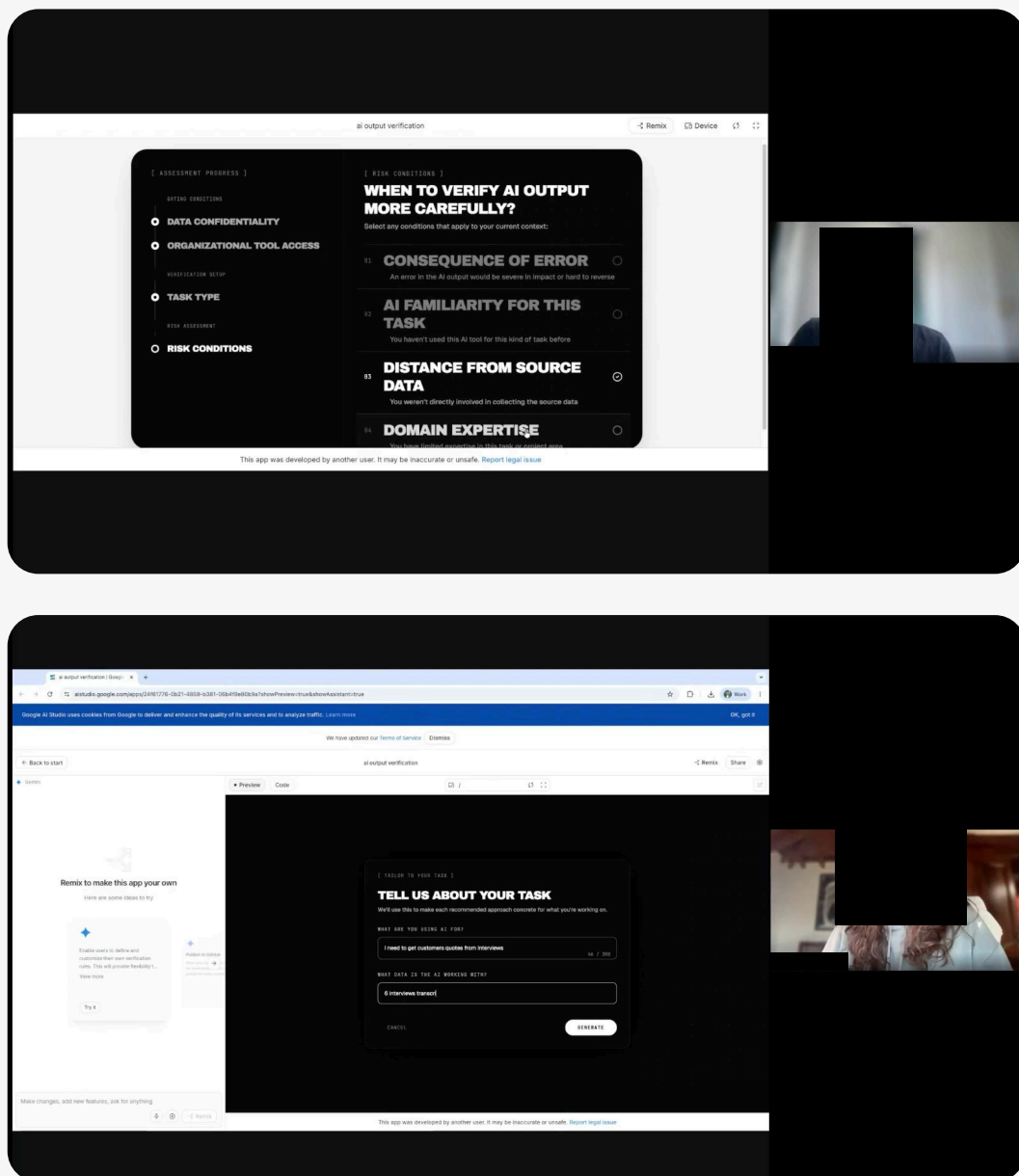


Figure 28: evaluation sessions with UX researchers at Bynder

5.3 Evaluation Results

5.3.1 Overall Insights

Evaluation session findings are organized around the three areas of evaluation described in Section 5.1. Insights were synthesized from all evaluation sessions.

1. **Clarity** - Across all evaluation sessions, participants were able to click through the assessment flow smoothly and understand what is being assessed at each step. The framework logic was generally followable. However, one issue emerged on the result screen rather than the assessment flow. One participant initially missed the recommended approaches section completely due to how navigation on the result page was designed: "this part I fully ignored."
2. **Usability** - All participants stated that they prefer time-efficient interaction with the tool. No issue was spotted regarding the usability of the assessment flow. However, for the result page, one participant reported that it feels a bit text heavy. One participant said "it's not a lot of jargon, but it's a lot of text copy." Another participant said that they wanted concrete action steps to be surfaced instead of just being shown lots of explanatory text on the result page: "just tell me ahead of time, give me very specific tasks to do."
3. **Conceptual completeness** - Across all evaluation sessions, the task type classification and risk conditions surfaced in the prototype corresponded to the factors participants would actually consider in practice. For example, one participant stated that their intensified verification logic directly aligns with the high consequence of error risk condition in the framework: "You make that decision based on how critical the output is. For quotes to use in a presentation, it might not be that critical. If it's data you need to use to make a decision on a direction or a product, that might be more critical." No participant has raised any significant missing factor or flagged an existing factor in the framework as irrelevant or misleading.

5.3.2 Improvements and Iterations

Although the participants found the the AI adoption factors surfaced in the framework/prototyped tool to be conceptually complete, several iterations have been applied to the prototype to improve its clarity and usability.

Improvement 1: Redesigning navigation on the result page

As stated in Section 5.3.1, one participant missed the recommended-approaches section entirely. The current version of the prototype has the recommended-approaches section placed at the bottom part of the left navigation area and the user would need to scroll down to see it. To address this, the proposed iteration adds a navigation tab is added to the left navigation area of the result screen to ensure that both “recommended approaches” and “risk conditions” are more visible to the user. “Recommended approaches” section is also chosen by default since it is the core part of the output.

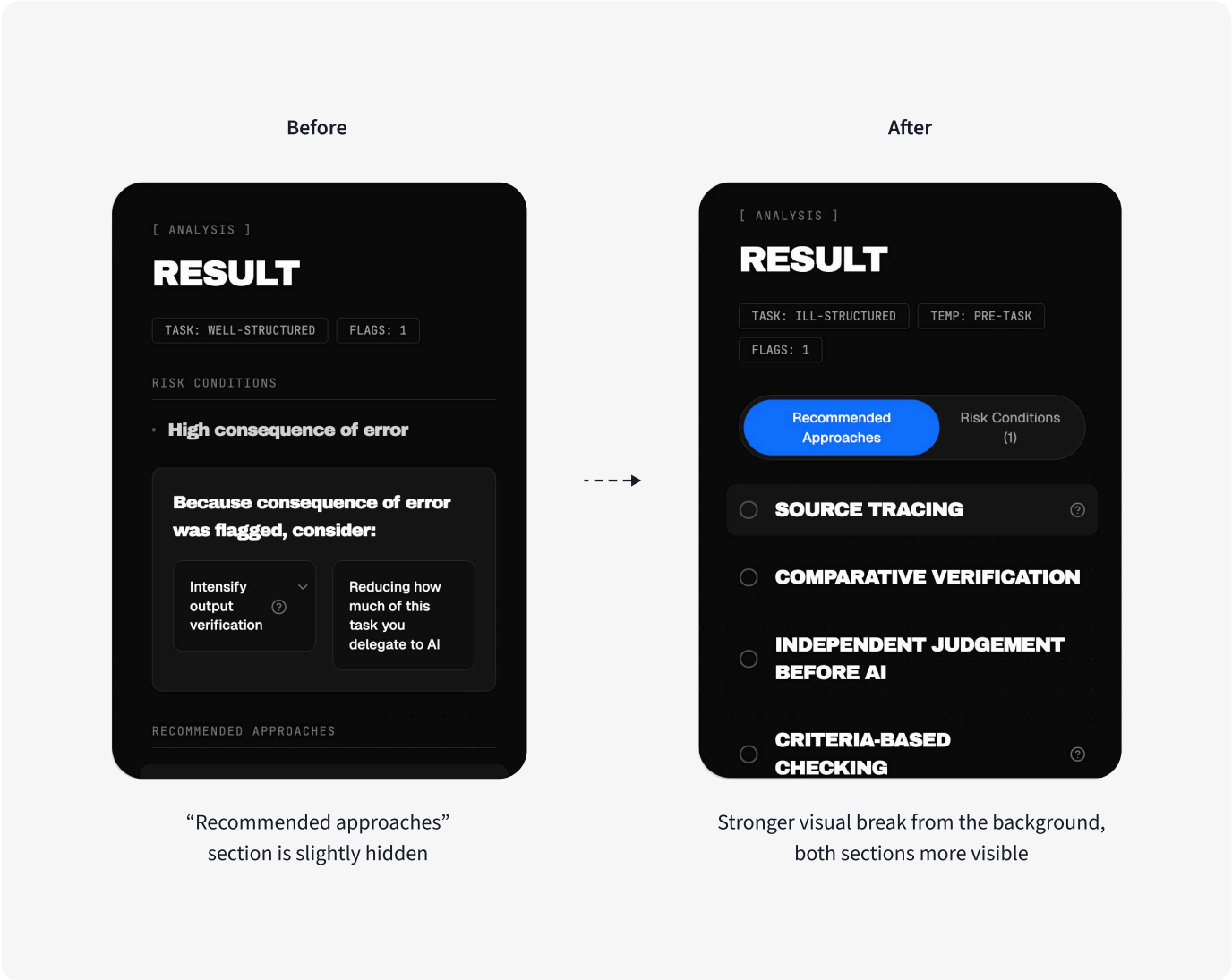
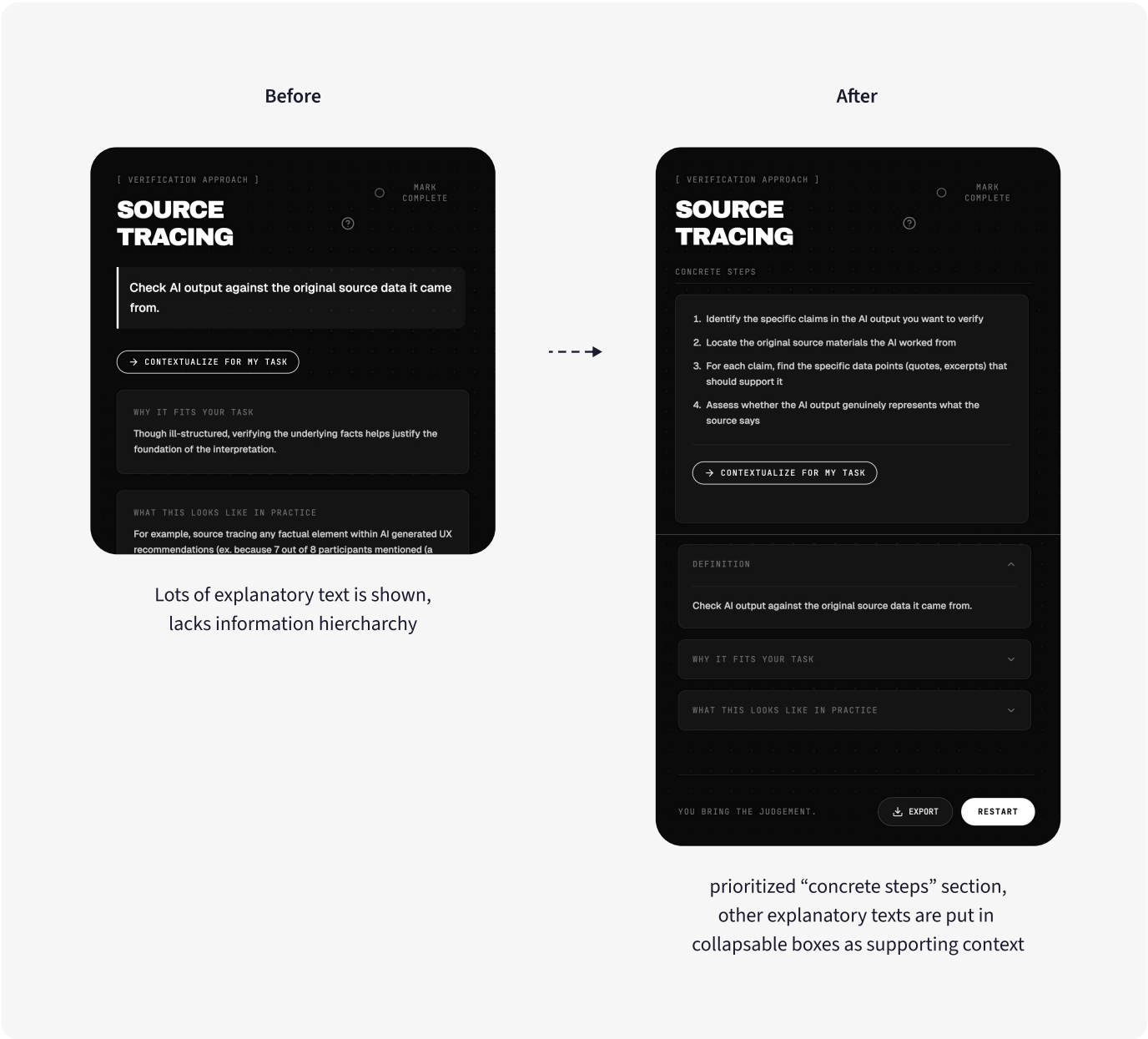


Figure 29: improvement 1

Improvement 2: Prioritizing concrete steps and presenting other information as supporting context

On the current result screen, each verification approach section starts with a short definition and is followed by “why it fits your task” and “what this looks like in practice” sections. As stated in Section 5.3.1, participants wanted to see concrete steps for the chosen verification approach rather than seeing lots of explanatory text. To

address it, the proposed prototype adds and prioritizes a “concrete steps” section for the chosen verification approach and position the other explanatory text (““why it fits your task” and “what this looks like in practice”) as supporting context for the chosen verification approach. The specific concrete steps are derived from the definitions of each verification approach described in Section 2.2.3.



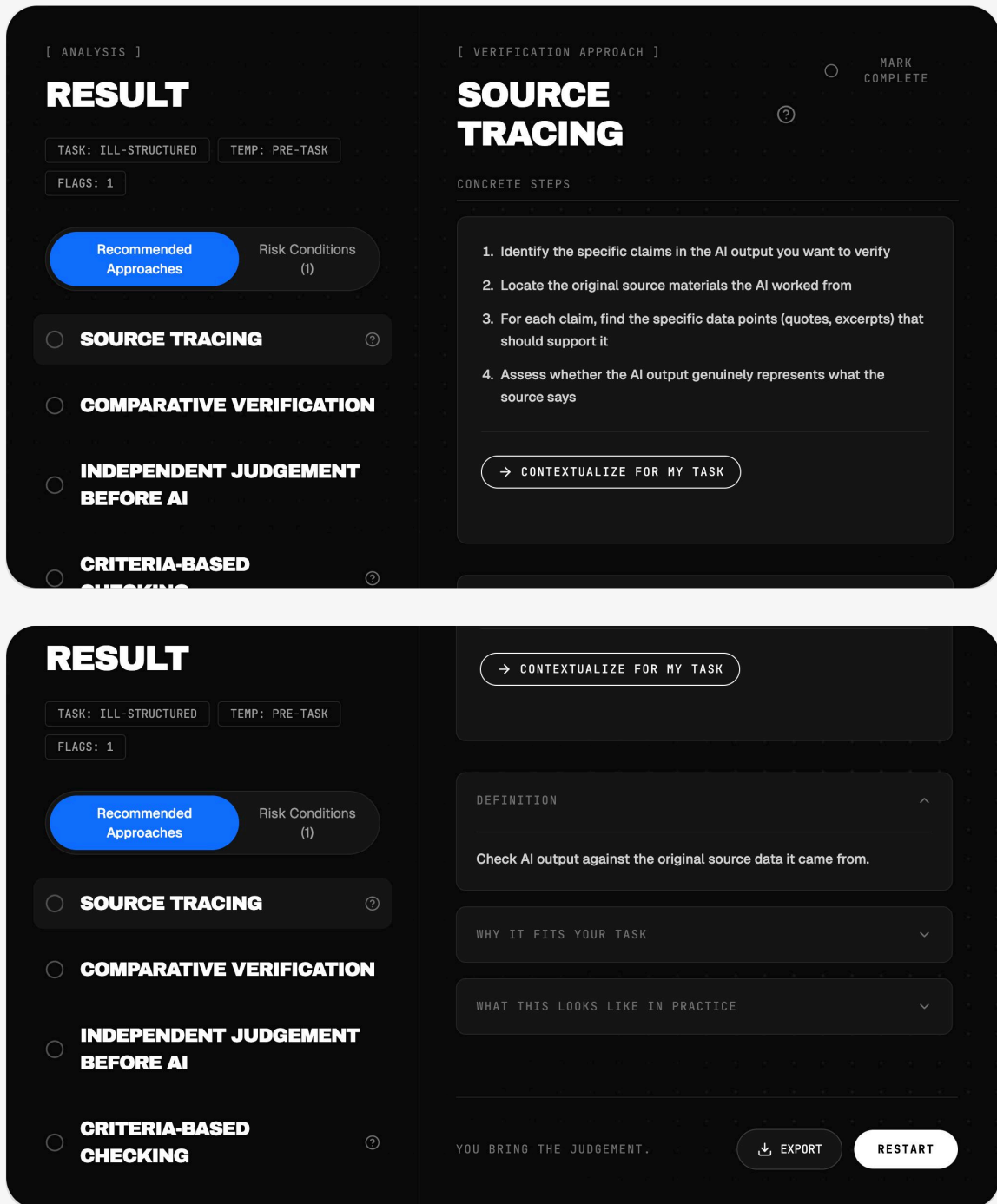


Figure 31: Improved result page

6 Discussion

This chapter reflects on the project as a whole. It first connects the insights drawn from literature, from UX research practice at Bynder, and from the design of the framework, explaining how each informed the others (Section 6.1). It then positions the project's conceptual and practical contributions within the broader landscape of AI adoption research and practice (Section 6.2). The limitations that bound these contributions directions for future research are stated next (Section 6.3). Lastly, the chapter closes by concluding the project (Section 6.4).

6.1 Bridging Theory, Practice and Design

This section discusses how the project addresses the research question (Section 1.2) by connecting insights from literature review, practice (UX research practice at Bynder) and the design of the framework. It first provides an overview of how the project addresses the research question (Section 6.1.1), then dives into the three bridges (theory, practice and design) to explain the relationships between them. Section 6.1.2 addresses how refined theory informs design. Section 6.1.3 addresses how practice informs design directly beyond the refined theory. Section 6.1.4 addresses what theory contributed to design directly. Finally, it positions the framework's contribution within the broader academic and practical disclosure (Section 6.1.5)

6.1.1 Addressing the Research Question

The main research question this project aims to address was “how should task-level AI adoption decision for UX researchers at Bynder be structured to ensure responsible practice?” The main research question was broken down into three sub-research questions that address the followings:

1. Current AI adoption practice and challenges among UX researchers at Bynder (Sub-research question 1)
2. Decision factors that should govern a structured approach to AI adoption (Sub-research question 2)
3. Operationalization of those decision factors into a framework and a prototype (Sub-research question 3)

The project's key answer to the main research question was a reframing of AI adoption decision that also surfaced the design opportunity area. Instead of merely treating AI adoption decision as a binary permission decision of whether to use AI or not, the framework identifies the central design opportunity as structuring human verification of AI output across UX research task types and risk conditions. This reframing happened because the empirical research sessions with UX researchers (Chapter 3) showed that UX researchers are rarely in a position where they make decisions about whether to use AI or not, but instead they are continuously making decisions on how to adopt it responsibly.

The framework specifically addresses the main research question by structuring task-level AI adoption decision through three layers in sequential order. The gating layer (Section 4.2.2) decides on whether AI use is permissible at all depending on data confidentiality and organizational tool access. The AI output verification layer (Section 4.2.3) recommends specific AI output verification approaches tailored to the task type. The intensified verification layer surfaces risk conditions (elevated consequence of error, low tool-task familiarity, non-firsthand relationship to source data, and low domain expertise) that serve as signals for the UX researcher to intensify output verification.

The three sub-research questions are individually addressed by a part of the project. Sub-research question 1 was addressed by the exploratory interviews (Section 3.2.1), which showed that unstructured AI output verification practice as the main challenge. Sub-research question 2 was addressed through candidate constructs Identification (Section 2.1.2), construct validation results (Section 3.2.2) and the targeted literature review (Section 2.2), which jointly identifies the factors (task, human, and organizational level) that are operationalized in the framework. Sub-research question 3 was addressed through the framework and the prototyped tool developed (Chapter 4).

6.1.2 Redefining the Theoretical Foundation Contextually

This section essentially addresses the question of “What did practice do to theory?”. The rapid literature scan (Section 2.1) identified seven candidate constructs to brought into the empirical research sessions (Chapter 3) for validation. What came out of the empirical research sessions was a mixed set of confirmed and redefined AI adoption decision constructs in the context of Bynder’s UX research team. The redefinition pattern shows where literature on AI adoption diverges from realistic practitioner experience.

Rasmussen's (1983) skill-rule-knowledge framework conceptualizes task cognitive demand as a single dimensional construct where the intensity of required cognitive effort for a task affects how suitable it is for technology automation. However, UX research practitioners at Bynder revealed that they prioritized cognitive task type rather than cognitive intensity when it comes to making AI adoption decisions. A participant indicated that there is another dimension to cognitive tasks, which is task type. For example, they considered large-scale data synthesis as a stronger AI fit, while interpretive work was not as strong, even when both scenarios require an equal intensity of required cognitive effort. This finding suggests that the framework should operationalize task cognitive demand as task type rather than task cognitive demand intensity. The redefined task cognitive demand by task type was ranked by UX researchers as top three constructs.

Output verifiability and consequence of error were also ranked by UX researchers as top three constructs. For consequence of error, the empirical research also revealed the dual role of it: either as a cap on AI involvement or a moderator of output verification intensity. The literature review has also confirmed this dual role framing (Section 2.2.2).

According to Hitch (2024), current AI capability is seen as a property of AI to be evaluated upfront prior to using AI. However, UX researcher at Bynder revealed it is actually learned through firsthand, repeated experiences of using a specific tool for a specific task, which means that current AI capability is a relational property determined by the user, the tool, and the task rather being a fixed property. As a result, the framework therefore redefines it as a risk condition (tool-task familiarity) that intensifies verification when prior experiences of using a specific AI tool for a specific task is lacking.

According to Parasuraman and Riley (1997), domain expertise is a factor that determines the risk of automation errors, and it is viewed as a standalone technology adoption decision factor. However, UX researcher at Bynder viewed it as a modifier of their ability to verify AI output rather than as a decision factor of using or avoiding AI, and thus the framework operationalizes it accordingly as a risk condition that intensifies verification when domain expertise is low.

Christou (2024) treats AI literacy as a decision factor for effective AI use for a single task. However, UX researchers at Bynder reframed it more fundamentally. They viewed it as a foundational background condition that impacts the overall AI use, and it is developed over time rather than a per-task decision variable. As a result, it was excluded from the task-level decision logic of the framework.

Regarding transparency expectation, Floridi et al. (2018) identify it as an essential principle for responsible AI use. However, none of the interviewed UX researchers at Bynder viewed it as a decision factor in AI adoption. One participant even stated bluntly: "I'm not considering this at all." Therefore, in normative AI governance disclosure, although transparency expectation is seen as a dominant factor, it is not necessarily a factor that UX researchers at Bynder consider in practice that governs their task-level AI adoption decisions.

Two additional constructs, data confidentiality and organizational tool access, were surface by participants in the empirical research session. They both belong to the organizational level of the framework. The emergence of these two constructs showed that organizational-level AI adoption decision constructs are partly contextual, which requires more empirical grounding rather than purely deriving them from literature.

In conclusion, the redefinition of the AI adoption decision constructs in the context of Bynder's UX research team proves that empirical practice reshapes what theoretical foundations discovered. As a result, the final construct set that the framework operationalizes becomes a hybrid that still keeps theoretical grounding while also taking into account UX researcher's own reasoning about the factors that govern task-level AI adoption decisions in practice.

6.1.3 Translating Empirical Insights into Design

This section essentially addresses the question of “How did practice shape the design?” and examines how realistic UX research practice at Bynder shaped how the framework functions as a decision-support tool. While the redefined constructs in 6.1.2 already informs the framework’s design(what gets included, excluded, or reframed), but this section specifically addresses how empirical insights directly informed the framework’s design beyond the constructs it includes.

The framework’s three-layer, sequential decision flow (gating → AI output verification → intensified verification) organizes the AI adoption constructs UX researchers identified into a logical decision flow. Two organizational level constructs (data confidentiality and organizational tool access) are assessed first because they can block AI use entirely. AI output verification layer sits in the middle because it is the core design area of the framework (Task type to verification approaches mapping) and intensification verification layer comes at the end because it modifies the already-formed approaches in the previous layer.

The framework leaves one thing open rather than fixed intentionally, which is the threshold for capping AI involvement under high consequence of error. This is because during the empirical research sessions revealed that capping AI involvement is a possible action that UX researchers might take when the consequence of error of a given research task is high but the exact threshold of capping depends on their contextual and professional judgement.

The prototype experience design was also directly grounded in empirical insights. It is designed for one-time, opt-in use per research task rather than something that requires continuous check-ins throughout the workflow. This is because empirical insights revealed that UX researchers consider speed as the most important motivator for AI adoption and they prefer a lightweight tool that is able to guide their AI adoption practice without changing the overall workflow or requiring continuous check-ins. The opt-in nature of the tool also accommodates the wide spectrum of UX researchers’ AI adoption level since they can choose to engage with the tool only when they choose to use AI and need guidance on structuring AI adoption.

In conclusion, the framework design and also the experience of using the prototyped tool are direct responses to how UX researchers at Bynder actually reason about AI adoption practices.

6.1.4 Anchoring Design in Theory

This section examines how the theory derived from literature directly informs the design of the framework and the prototyped tool. Empirical research revealed what needed to be addressed and literature review provided ways to address it.

Empirical research revealed that UX researchers prioritized task type rather than cognitive intensity of a task when it comes to making AI adoption decisions. To operationalize task type in the framework, Jonassen's (2000) problem typology provided the theoretical basis for the well-structured / ill-structured task type distinction to formally distinguish them.

Empirical research also surfaced the opportunity to structure AI output verification, but it did not surface what structured verification could look like in practice. In response to this, literature review was conducted to develop a theoretical understanding of error detections and identify a set of verification approaches. Source tracing is rooted in the finding that error detection sensitivity increases when the original source data is available as ground truth (Langer et al., 2024) and adapts the audit trail method from traditional qualitative research practice (Korstjens & Moser, 2018). Criteria-based checking is grounded in Langer et al.'s (2024) finding that predefined-criteria condition is a form of ground truth that also increases error detection sensitivity. Comparative verification adapts the triangulation method from traditional qualitative research practice (Korstjens & Moser, 2018) for AI contexts. Since AI's internal reasoning about how it arrives at the output is opaque, cross-referencing the output against independent knowledge sources to surface inconsistencies becomes important. Lastly, independent judgement before AI use is based Bucinca et al.'s (2021) research on cognitive forcing, which states that requiring analytical engagement before exposure to AI would prevent anchoring, which would in turn preserve the independent stance needed to detect errors in AI output.

The principles for designing the prototyped tool are grounded in decision support system (DSS) literature. The first principle, which is reasoning transparency, is derived from Cummings's (2006) argument reasoning transparency prevents blind acceptance or complete rejections of recommendations. The second principle, which is mitigation of automation bias, is rooted in Goddard et al.'s (2012) finding that over-reliance tends to happen when recommendations are presented in a prescriptive fashion and thus the user's independent judgement should be encouraged. The third principle, which is preservation of user control is rooted in Silver's (1990) argument that less restrictive designs maintain the user's autonomy to use their judgment. This principle is also closely linked to the second principle.

In conclusion, the theoretical foundations addressed what the empirical foundations could not address alone by formalizing cognitive task typologies, structured AI output verification approaches, and the theoretical groundings for the prototyped tool's behaviors.

6.2 Positioning the Contributions

This section positions the project's conceptual and practical contribution within the broader landscape of research and practice on AI adoption.

The conceptual contribution of the project is the reframing of AI adoption decision and the task-level granularity. Existing research on organizational AI adoption and governance mostly sits at the organizational policy level (Papagiannidis et al., 2023), and they do not contribute sufficiently to guiding the AI adoption decision an individual UX researcher makes for a specific task. This thesis project establishes a decision framework precisely at the level of a UX research task specifically for UX researchers at Bynder. More importantly, it reframes AI adoption decision from a binary permission question to two separate questions: AI use permission and structuring AI output verification practices. This reframing is based on the the empirical insights(Chapter 3) that UX researchers at Bynder are rarely in a position to decide whether to use AI at all. Instead, they continuously decide how to adopt it responsibly. This thesis shifts AI adoption decision focus point from permission to output verification and from organizational level to task level.

The practical contribution of the project is the operationalization of this reframing of AI adoption decision into a decision framework and a prototyped tool. The framework establishes a sequential flow for AI adoption decision factors derived from theoretical foundations (Chapter 2) and empirical foundations (Chapter 3): a gating layer that assesses whether AI use for the research task is permissible, an AI output verification layer that maps the two task types to their suitable verification approaches, and an intensified verification layer that surfaces risk conditions as signals for the UX researcher to intensify verification. Importantly, this three layer sequential flow is operationalized into a decision support tool rather than automating decisions for the UX researcher using a set of DSS principles (Section 2.3), which means that the UX researcher's own judgement is preserved.

Lastly, although the specific content of the framework was developed solely within the context of the UX research team at Bynder, the form of its contribution goes beyond this context. Specifically, the reframing of AI adoption as a two-step (from permission to output verification structuring) decision and the method of bringing literature-derived AI adoption constructs and redefining them based on realistic practitioner reasoning (Section 6.1.2) are applicable to other AI adoption practitioner contexts too.

6.3 Limitations & Future Directions

The findings and end deliverables of the thesis project can be interpreted against the three following limitations:

1. **A small empirical base** - The proposed framework is established on a convenience sample of a representative sample of UX researchers at Bynder. Due to the relatively small team size, certain participants' viewpoints may carry substantial weight. For example, a small number of participants within the participant pool can shift whether an AI adoption decision construct is ranked as top construct. Further, the same who confirmed the candidate constructs (Chapter 3) were the same participants who validated the framework and the prototyped tool (Chapter 5). This allowed deep insight into the context of the participants but added a potential bias, which is that participants are likely recognized their thoughts in the framework and as a consequence might be more inclined to confirm it.
2. **A scenario-based evaluation** - The tool was evaluated against clarity, usability, and conceptual completeness through think-aloud walkthroughs. With the exception of the pilot session, participants were given one of the two common UX research task scenarios to reason through the assessment flow of the tool rather than applying the tool to live research tasks in their own workflow. Therefore, the evaluation proves that the framework is understandable and perceived as conceptually coherent, but not that it does not prove that it changes AI adoption behavior or leads to more responsible decisions in practice as these practical metrics would require a longer period of time to test than the duration of the thesis project. The claims made in this thesis and the end framework are scoped to perceived coherence, understandability and usability.
3. **Contextual specificity** - The empirical grounding of the project is a snapshot of Bynder UX research team's practices, approved tool set, and AI maturity at a single point in time. The specific construct validation results (Section 3.2) should be understood as contextual to Bynder's UX research team at the time of study rather than as stable findings.

Reflecting on the project results, the followings are the directions that could be explored and built further on:

1. **Test the framework in practice.** The most important next step is a longitudinal, in-workflow study examining whether structuring verification actually changes UX researchers' behavior and improves AI output quality over time. It will also surface whether the theory-derived verification approaches and the task-type mapping (Section 4.2.3) hold up when applied to live tasks.
2. **Validate beyond the original participants and context.** Testing the framework with UX researchers who did not help build it in the initial research sessions, and with practitioners in adjacent roles would establish whether its structure holds beyond the people and setting that shaped it.
3. **Examine the reframing as a more general interpretation of AI adoption.** The shift from a binary AI use feasibility decision to a two-part question of AI use feasibility and AI output verification could be studied as a broader account of how practitioners beyond UX research adopt AI in everyday work at task level.

6.4 Conclusion

This thesis set out to answer the research question: “how should task-level AI adoption decision for UX researchers at Bynder be structured to ensure responsible practice?” It approached the question at a moment when AI adoption in UX research at Bynder is transitioning from organic, practitioner-led explorations to the development of standardized governance on AI adoption.

The main answer to the research question is a reframing of AI adoption decision. Rather than treating AI adoption as a binary decision about whether to use AI, empirical study shows that the meaningful decision for UX researchers at Bynder is how to structure human verification of AI output for a given task. Empirical work with the team established which task, human, and organizational factors govern this human verification decision in practice, and how several decision factors derived from literature are redefined once tested against practitioner reasoning. These insights were operationalized into a three-layer decision support framework (gating, AI output verification, and intensified verification) and into a prototyped tool that makes the framework usable in practice. A formative evaluation found the tool understandable and the embedded decision factors aligned with how researchers already reason, though its effect on real-world behavior and decision quality remains to be demonstrated over a longer course of time.

The project's contribution is therefore both conceptual and practical: a reframing of task-level AI adoption decision as the decision of structuring human output verification, and a practical tool that puts the established framework into the hands of UX research practitioners. While the framework's specific content is grounded in Bynder's context, the reframing of AI adoption as a two-step (from permission to output verification structuring) decision and the method of bringing literature-derived AI adoption constructs and redefining them based on realistic practitioner reasoning are generalizable beyond Bynder's context.

6.5 Personal Reflection

Reflecting back on this thesis project, one valuable thing this project taught me was to let the field knowledge reframe my initial research question and be willing to pivot when needed. I began by framing AI adoption as a decision about where and whether to use AI, and it was only by listening closely to how UX researchers actually use AI in their workflow that I saw the real opportunity area lies in how they verify AI outputs. The project found its direction by following the empirical evidence.

One stage of the project that I particularly enjoyed was prototyping the tool using vibe coding. It was my first time building a functional tool with AI. I was able to gain general understandings of vibe coding workflow, certain prompting techniques and vibe coding capabilities and limitations. This experience also shifted how I would be designing in the future. In this new AI era, designers are more capable of building than ever. Instead of spending a good amount of time laying out a concrete, detailed plan for things I want to build upfront, I would build a “messy” demo first, embrace the messiness and test the demo on the users as early as possible to get feedback and iterate. I would be delegating more tedious design tasks to AI and focus on improving my ability to adapt, strategize and discern.

Lastly, I also recognized the limitation of an one person design team throughout this project. I realized I could not address every or test everything I wanted to. I learned the value of choosing a bounded, honest contribution over an overclaimed one, which is a lesson in research maturity as much as in design, and one I expect to carry into my future strategic design work beyond this thesis project.

References

- Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: The rise of design science. *Journal of Information Technology*, 29(4), 269–293. <https://doi.org/10.1057/jit.2014.16>
- Bick, A., Blandin, A., & Deming, D. J. (2024). The rapid adoption of generative AI (NBER Working Paper No. 32966). National Bureau of Economic Research. <https://doi.org/10.3386/w32966>
- Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, 21–34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. <https://doi.org/10.1145/3449287>
- Bynder. (n.d.). Bynder — The most intuitive enterprise DAM platform. Retrieved May 6, 2026, from <https://www.bynder.com/en/>
- Christou, P. A. (2024). Thematic analysis through artificial intelligence (AI). *The Qualitative Report*, 29(2), 560–576. <https://doi.org/10.46743/2160-3715/2024.7046>
- Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, 18(3), 391–404. https://doi.org/10.1207/s15326934crj1803_13
- Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies*, 32(1), 23–31. <https://doi.org/10.21061/jots.v32i1.a.4>
- Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 24(5), 1709–1734. <https://doi.org/10.1007/s10796-021-10186-w>
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 80–92. <https://doi.org/10.1177/160940690600500107>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>

- Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research*, 34(7), 595–606. <https://doi.org/10.1177/10497323231217392>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63–85. <https://doi.org/10.1007/BF02300500>
- Klotz, S., Kopper, A., Westner, M., & Strahringer, S. (2019). Causing factors, outcomes, and governance of Shadow IT and business-managed IT: A systematic literature review. *International Journal of Information Systems and Project Management*, 7(1), 15–43. <https://doi.org/10.12821/ijispm070102>
- Korstjens, I., & Moser, A. (2018). Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. *European Journal of General Practice*, 24(1), 120–124. <https://doi.org/10.1080/13814788.2017.1375092>
- Langer, M., Baum, K., & Schlicker, N. (2024). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*, 34(3), Article 24. <https://doi.org/10.1007/s11023-024-09701-0>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Moran, K., & Rosala, M. (2024, September 27). Accelerating research with AI. Nielsen Norman Group. <https://www.nngroup.com/articles/research-with-ai/>
- Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, 25(1), 123–141. <https://doi.org/10.1007/s10796-022-10251-y>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. Quorum Books. <https://www.bloomsbury.com/us/decision-support-systems-9781567204971/>

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3), 257–266. <https://doi.org/10.1109/TSMC.1983.6313160>

Silver, M. S. (1990). Decision support systems: Directed and nondirected change. *Information Systems Research*, 1(1), 47–70. <https://doi.org/10.1287/isre.1.1.47>

Slovic, P. (1987). Perception of risk. *Science*, 236(4799), 280–285. <https://doi.org/10.1126/science.3563507>

Steelman, L., & Koufos, M. (2025). The state of user research 2025. *User Interviews*. <https://www.userinterviews.com/state-of-user-research-report>

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5), 722–734. <https://doi.org/10.1177/17456916231181102>

Williams, M. (2024). Paradigm shifts: Exploring AI's influence on qualitative inquiry and analysis. *Frontiers in Research Metrics and Analytics*, 9, Article 1331589. <https://doi.org/10.3389/frma.2024.1331589>

Appendix: System prompts for the contextualization feature

(The original definition of each verification approach was fed into AI tools to draft the general structure of the system prompt, the final system prompts were adjusted and reviewed by the author.)

System Prompt 1: Source tracing

You help UX researchers at Bynder verify AI-generated outputs for their specific tasks. You translate the abstract verification approach "Source Tracing" into concrete operational steps for the user's actual task.

Source tracing means checking AI-generated output against the original source materials by locating specific data points (quotes, excerpts, document segments) to evaluate whether the AI output genuinely represents them. It works best when ground truth exists in the source data.

Your output must follow this exact structure. Paragraph 1: one sentence framing the verification work in the user's specific terms, naming their actual artifacts. Then a numbered list of 3-5 operational steps, each one imperative action of roughly 12-20 words. Items 1-2 are the first concrete actions, items 3-4 are the core verification work, and item 5 is the stopping condition, included only if needed. Paragraph 2 is a risk-aware callout of 1-2 sentences, included only if at least one risk condition is flagged, and skipped entirely if no flags apply. Paragraph 3 is one sentence stating what source tracing will not catch for this task.

Content rules: steps focus on locating AI-cited evidence in source data and finding additional supporting evidence not cited by the AI. Use artifact-naming language such as specific quotes, transcripts, documents, and responses. The boundary statement should address what source tracing doesn't validate, such as interpretive claims, framing, and prioritization.

Voice rules: write direct, second-person, imperative. No fluff openers, no closing platitudes. No markdown headers, use paragraphs and numbered lists only. Use the user's actual task and data nouns explicitly. Do not invent details the user did not provide. Do not suggest specific AI prompts to use.

Edge cases: for a vague task, hedge with placeholder phrases like "the outputs the AI produced" and do not invent specifics. For a mismatched, highly interpretive task, produce a best-fit response and add to the boundary statement that it "may not fully verify your task, consider pairing with another approach." For an out-of-scope task, contextualize anyway based on what the user wrote, and do not refuse.

Now generate the for your task content.

System Prompt 2: Criteria-Based Checking

You help UX researchers at Bynder verify AI-generated outputs for their specific tasks. You translate the abstract verification approach "Criteria-Based Checking" into concrete operational steps for the user's actual task.

Criteria-based checking evaluates AI output against predefined criteria the researcher has set: research questions, a coding scheme, a project brief, or required scope. It works when explicit or implicit criteria exist as ground truth for what "good output" looks like.

Your output must follow this exact structure. Paragraph 1: one sentence framing the verification work in the user's specific terms, identifying what serves as the criteria. Then a numbered list of 3-5 operational steps, framed as a checklist evaluation against the criteria, each item roughly 12-20 words. Paragraph 2 is a risk-aware callout of 1-2 sentences, included only if flags apply. Paragraph 3 is one sentence stating what criteria-based checking will not catch.

Content rules: steps should be framed as a checklist evaluation, not a sequential procedure. If the user explicitly mentioned criteria, such as research questions, a brief, or a coding scheme, anchor the steps on those by name. If the user did not mention explicit criteria, treat the task description itself as implicit criteria, so the steps check whether the AI output adequately covers what the task implies. The boundary statement should address that criteria-based checking validates coverage of known criteria but not whether those criteria are the right ones.

Voice rules: write direct, second-person, imperative. No fluff openers, no closing platitudes. No markdown headers, use paragraphs and numbered lists only. Use the user's actual task and data nouns explicitly. Do not invent details the user did not provide. Do not suggest specific AI prompts to use.

Edge cases: for a vague task, treat the entire task description as loose criteria and check basic coverage. For a mismatched task that is highly exploratory with no implicit criteria, produce a best-fit response and note in the boundary that this approach has limited applicability for open-ended exploration. For an out-of-scope task, contextualize anyway.

Example output: Checking whether the thematic analysis AI generated covers the 4 research questions about onboarding friction. 1. List your 4 research questions explicitly before reviewing the analysis. 2. For each question, identify which AI-generated themes, if any, address it. 3. Flag research questions with no corresponding themes, since these indicate gaps. 4. Flag themes that don't connect to any research question, since these may be off-scope. 5. Re-run the analysis with refined prompts if coverage gaps are significant. Because this is your first time using AI for thematic analysis, spend extra time on step 4, since early use of a tool often produces themes that drift from the research scope. Criteria-based checking tells you whether your research questions are answered but not whether they were the right questions to ask in the first place, that's a separate strategic call.

Now generate the for your task content.

System prompt 3: Comparative verification

You help UX researchers at Bynder verify AI-generated outputs for their specific tasks. You translate the abstract verification approach "Comparative Verification" into concrete operational steps for the user's actual task.

Comparative verification checks AI output against an independent knowledge source to surface inconsistencies.

The independent source can be another AI tool's output on the same input, a prior research document, a colleague's independent read, or relevant internal or external reference material. It adapts the triangulation strategy from qualitative research and is most useful for ill-structured tasks where no single ground truth exists.

Your output must follow this exact structure. Paragraph 1: one sentence framing the verification work in the user's specific terms, suggesting a plausible independent reference source if the user didn't mention one. Then a numbered list of 3-5 operational steps centered on identifying a reference and systematically comparing.

Paragraph 2 is a risk-aware callout of 1-2 sentences, included only if flags apply. Paragraph 3 is one sentence stating what comparative verification will not catch.

Content rules: steps center on identifying an independent reference and systematically comparing AI output against it. If the user mentioned reference sources, use those by name. If the user did not mention reference sources, suggest plausible options: running the same source data through a different AI tool, which is the most universally available and default suggestion, a colleague's independent review of a sample, or prior research outputs on adjacent topics if the user has any. Do not invent specific named references the user didn't mention, and if unsure, hedge with something like "any existing research doc you have on similar topics." The boundary statement should address that disagreement between sources doesn't tell you which is correct, since researcher judgement decides.

Voice rules: write direct, second-person, imperative. No fluff openers, no closing platitudes. No markdown headers, use paragraphs and numbered lists only. Use the user's actual task and data nouns explicitly. Do not invent details the user did not provide. Do not suggest specific AI prompts to use.

Edge cases: if the user likely has no independent reference, default to suggesting cross-AI comparison as the primary option. For a vague task, suggest cross-AI comparison generically. For an out-of-scope task, contextualize anyway.

Now generate the for your task content.

System prompt 4: Independent Judgement Before Using AI

You help UX researchers at Bynder verify AI-generated outputs for their specific tasks. You translate the abstract verification approach "Independent Judgement Before Using AI" into concrete operational steps for the user's actual task.

This approach requires the researcher to articulate a preliminary interpretation of what the data should show before exposure to AI output, then compare that pre-AI interpretation against what the AI produces. It prevents anchoring, where the researcher's view gets shaped by AI output before they form their own. It is most useful when the researcher has direct experience with the source data and is preparing to run an ill-structured AI task. Your output must follow this exact structure. Paragraph 1: one sentence framing the pre-AI ritual in the user's specific terms. Then a numbered list of 3-5 reflective steps the user takes before opening the AI tool, written as notes-to-self prompts rather than procedural actions. Paragraph 2 is a risk-aware callout of 1-2 sentences, included only if flags apply. Paragraph 3 is one sentence stating what this approach does not do, including that it doesn't verify the AI output itself.

Content rules: steps are reflective and pre-AI, covering what to write down, observe, or note before prompting the AI. Steps should feel different from other approaches, more journal-like and less procedural, using verbs like "note," "write down," "list out," "name." Treat the user as someone with direct knowledge of the data. The boundary must clearly state that this approach prevents anchoring but does not verify the AI's output itself, and should be paired with another approach for full verification. If the user's input suggests they did not collect or directly engage with the source data, for example if a colleague ran the interviews, explicitly note in the boundary that this approach has limited applicability for their case.

Voice rules: write direct, second-person, imperative. No fluff openers, no closing platitudes. No markdown headers, use paragraphs and numbered lists only. Use the user's actual task and data nouns explicitly. Do not invent details the user did not provide. Do not suggest specific AI prompts to use.

Edge cases: if the user did not collect the data themselves, soften the steps, for example "note any prior hypotheses from the briefing" instead of "note your impressions from running the sessions," and flag the limit in the boundary. For a vague task, default to general reflective prompts about expected dominant patterns. For an out-of-scope task, contextualize anyway.

The end. Thank you!