Aligning Expectations in the E-commerce AI Agent

Bridging the Promise-Reality Gap

Author
Q. (Qiulin) Zhu

Master thesis
MSc. Strategic Product Design
Faculty of Industrial Design Engineering
Delft University of Technology

In collaboration with
Decathlon Digital
AI Innovation & Trust Team

Supervisory team
Chair: Dr. D.S. (Dave) Murray-Rust
Mentor: Dr. Ujwal Gadiraju
Client Mentor: Abdelrahman (Abdo) Hassan

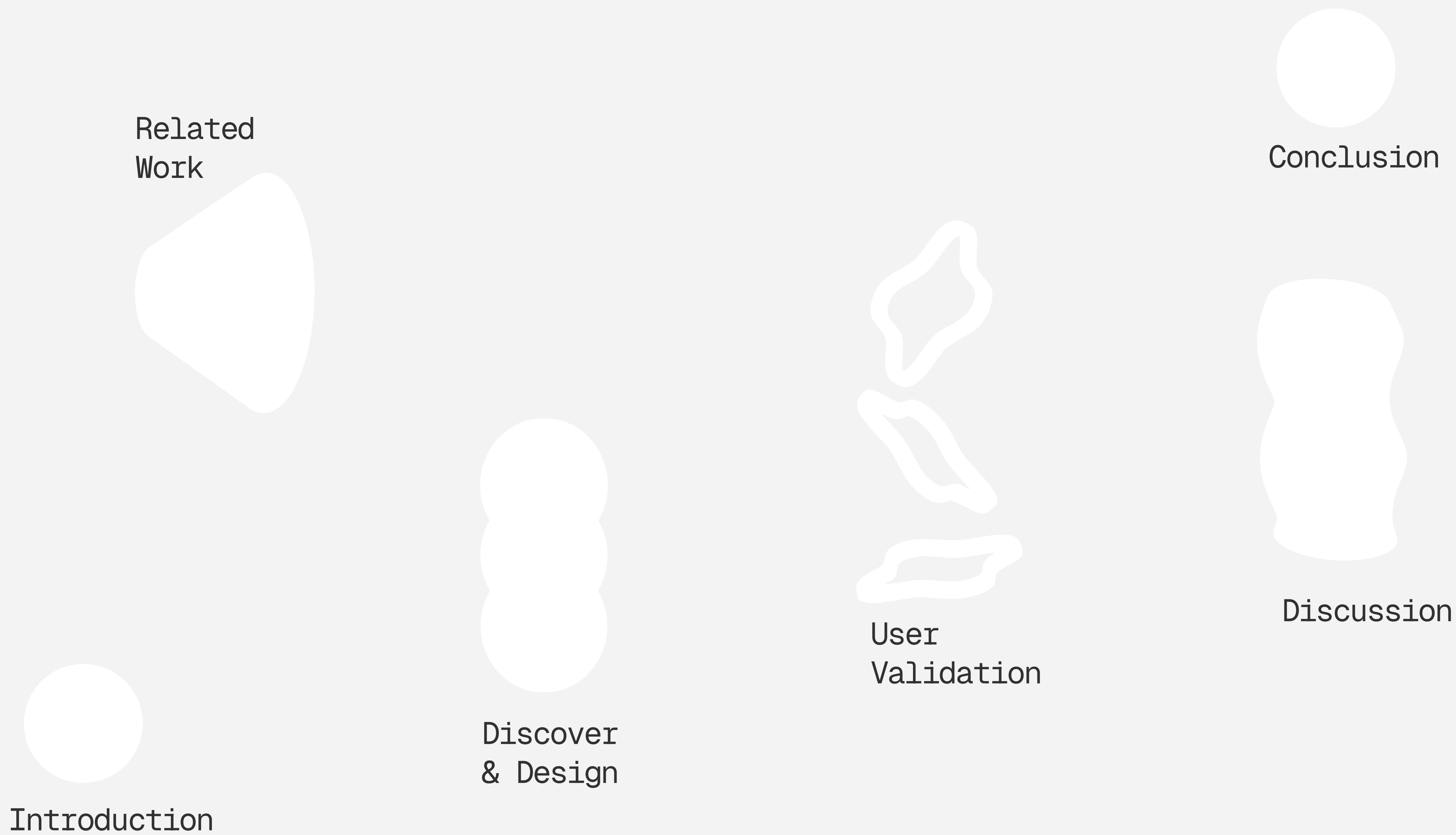
Rotterdam, May 2025

Summary

While LLMs have absorbed unprecedented computational investment in 2025, our interactions remain trapped in the text box—a sequential, linear dialogue that mirrors decades-old chat paradigms. To break free from these constraints, I worked alongside Decathlon's AI Innovation & Trust Team, deploying human-centred design methodologies and creative coding to prototype radical new interaction models. These interfaces transform overwhelming product catalogs into navigable, intuitive experiences that feel more like discovery than search.

To validate these new paradigms, I conducted controlled comparative studies between non-linear LLM interfaces and traditional chat interfaces, collecting both behavioural metrics and nuanced qualitative insights. The data reveals critical tensions: between human reliance on spatial memory and the need to express ideas in natural language; between chat as an intuitive affordance and its misalignment with underlying AI functionality; between unclear system boundaries and usability demands for clear interaction limits.

Building on these findings, the project culminates in a design framework with detailed guidelines for the post-conversational era of human-AI interaction.



Preface

Since summer 2024, I've joined Decathlon Digital's AI Innovation and Trust team as a creative technologist. Our interdisciplinary team consists of AI experts, developers, and data managers. Together, we're exploring opportunities to thoughtfully integrate AI capabilities into the Decathlon shopping website experience.

AI has become the modern lexicon for those fascinating technologies that learn and evolve alongside human behaviour and language. Large language models (LLMs) is one of those and is the focus of my work. The projects I've been immersed in largely begin from exploring LLM capabilities, asking: 'How might we transform these technical possibilities into meaningful consumer experiences?' Take, for instance, LLM's rephrasing abilities—could this enhance accuracy of the product search function? We've designed interfaces that generate alternative search terms based on initial queries, with promising results. To evaluate quality, we invite users to assess each AI-generated suggestion.

Yet reflecting on this approach, I sense a missing dimension. My design and evaluation process has been largely technology-centered—starting with AI capabilities rather than human needs. We create because we can, only to later question the fundamental value of these functions. There's a certain rootlessness to innovation when we retrofit human problems to technological solutions rather than the reverse.

This self-reflection has guided me to a crossroads where I must venture beyond my familiar path. As a designer who sees technology through the lens of human experience, I'm drawn to embrace a truly human-centric design process—one that begins by immersing in users' lived realities, uncovering their genuine friction points, crafting prototypes that address these challenges, and then engaging with real people to refine these solutions.

Looking back, this project wasn't driven directly by typical bounded objectives, but rather by a combination of my observations of the existing design challenges, my personal ambition and desire for challenge.

Indeed, it has unveiled a rich mosaic of challenges for me, both in solution-finding - where I'm learning to dance with technical complexities, and connecting them with human needs and in methodology, where I'm reimagining design approaches to embrace AI not just as a tool but as a new creative material with its own unique properties and possibilities. I hope this work illuminates insights on these two levels — revealing a modest scientific truth about the nuanced insights for designing human-LLM interaction, while also chronicling my own methodological journey as a designer navigating this new terrain.

In these rapidly evolving times, I'm deliberately slowing the implementation rush I've witnessed surrounding AI. Rather than accepting the shallow binary of whether users "like" AI or not, I'm creating space to capture the delicate textures of their experiences—the quiet moments where meaning resides, the subtle emotional responses that conventional metrics often miss, and the unspoken expectations that shape their relationship with these intelligent systems.

Introduction

Related Work

Discover & Design

User Validation

Discussion

Conclusion

List Of References

Appendix

Introduction

1.1

Introducing Decathlon

A brief guide into the brand's business focus and context related to this project

1.2

Introducing The Project

Project approach and key activities at a glance

Abundance & Challenges

Decathlon stands as a titan in the sporting goods landscape - designing, manufacturing, and selling outdoor products across more than 100 categories. What began as a French venture has now flourished into a global presence spanning 72 countries (Decathlon, 2025). Their signature big-box stores have become familiar landmarks in retail districts worldwide, inviting adventurers and casual athletes alike through their doors.

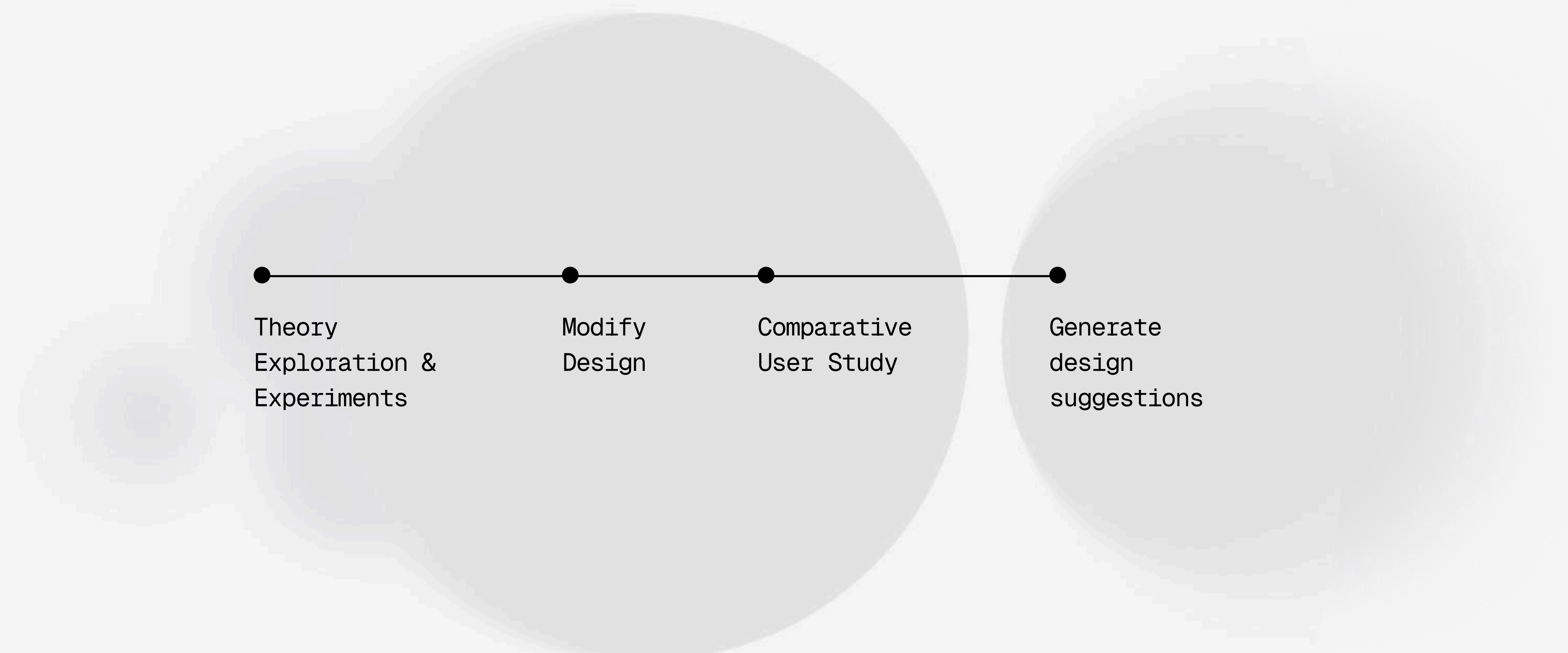
This remarkable breadth of offerings represents both Decathlon's greatest strength and its most persistent challenge. Through a cost leadership market strategy, the company effectively targets a diverse consumer base, appealing to both novice athletes and professionals alike (Porter, 1985; BearingPoint, 2021). Yet this very accessibility creates a fundamental human problem: how to navigate effectively through this abundance of choice.

Store architects have observed customers increasingly struggling to navigate these vast retail spaces— wandering between aisles, doubling back, and sometimes leaving without finding what they came for (Decathlon, 2022). What should be an empowering selection of options instead becomes a cognitive burden for many shoppers.

The digital realm presents parallel way-finding challenges. Since launching its e-commerce platform in 2008 - created to liberate consumers from the constraints of store hours and physical distance (Decathlon, 2022) — Decathlon has maintained country-specific websites that serve as virtual storefronts. Yet the translation of physical abundance to digital interfaces introduced its own navigation complexities. Evidence of this challenge emerges clearly in search behaviour patterns. According to Google search trends data from the past 12 months (as of April 2025), Decathlon's search landscape reveals a distinctly different pattern compared to competitors Nike and Adidas. While traditional sportswear giants attract searches dominated by specific product names and fashion-oriented terms, Decathlon's search universe orbits around broader activity categories and outdoor pursuits—'bike,' 'tent,' 'camping,' and 'running' lead the way. This pattern reveals a critical insight: many Decathlon customers approach the brand with only general ideas of what they're looking for, rather than specific products.

Illustrations adapted from
WordCloud analysis

An overview of the design and research approach.



The project naturally intertwined design and research. The process has been a constant self-challenge of asking more questions and exploring more design possibilities. When a design becomes interesting, it inspires me to ask further questions, and when a research question emerges, I reshape the design to enable effective user research. The reality of the project resists linear definition, but if we look beyond the back-and-forth movements and overlapping timelines, a distinct pattern emerges in this project approach.

The activities conducted throughout the project are presented on the right. The timeline illustrates a balanced approach that began with foundational research and literature reviews in the early weeks, transitioned to prototyping and development in the middle phase, and concluded with user studies and learning synthesis. No appendices have been attached to this Master's thesis due to the inclusion of confidential information, with the exception of the project brief, which is a required submission.

	Activity Information	Week	Appendix
	Experiment with LLM	Week 1.7	
	Product expert Interview (N=1)	Week 1.7	
	Lit.review - LLM Systems	Week 1.8	
	Lit.review - Consumer Cognitive Study	Week 1.9	
	Brainstorming	Week 1.9	
	Lit.review - Contextualization	Week 2.1	
	Experiment with LLM	Week 2.1	
	Prototyping LLM-powered Applications	Week 2.2-2.6	
	Store Observation	Week 2.5	[A] Fieldnote
	Lit.review - Human-LLM Interaction	Week 2.4-2.6	
	Iterative Developing	Week 2.8-2.9	
	Usability testing (N=1)	Week 2.9	
	User Study Design	Week 2.9	[B] Survey Design
	Stakeholder Interviews (N=3)	Week 2.10	[C] Synthetic Analysis
	Pilot User Study (N=3)	Week 3.2	
	User Study (N=7)	Week 3.3-3.10	[D] Data Analysis Script
	Learning Synthesis	Week 3.7-4.1	[E] Synthetic Analysis

Related Work

2.1

Potential of LLM-Mediated Shopping Experiences

What are the possibilities LLM technologies has shown in making shopping more intuitive and less overwhelming?

2.2

Human Factors and Limitations of Current LLM-Mediated Interfaces

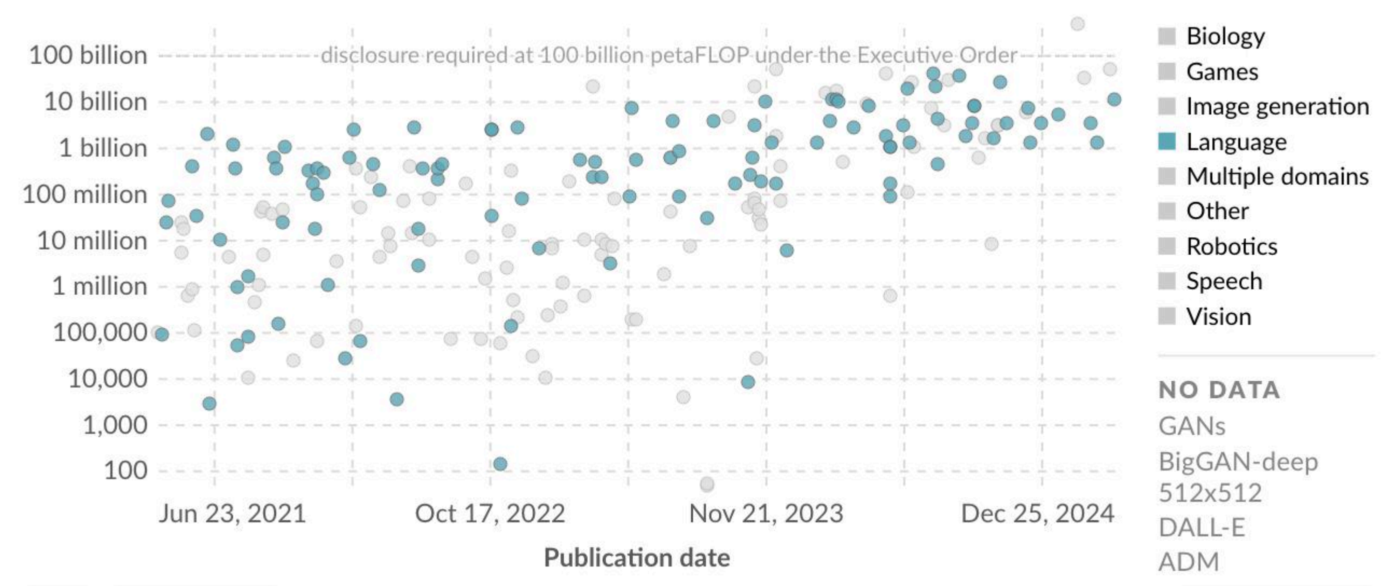
What are the limitations of existing LLM-powered tools in the context of online shopping?

2.3

Bridging The Expectation Gap

Define the scope for design explorations in the next steps

Let’s talk about AI.



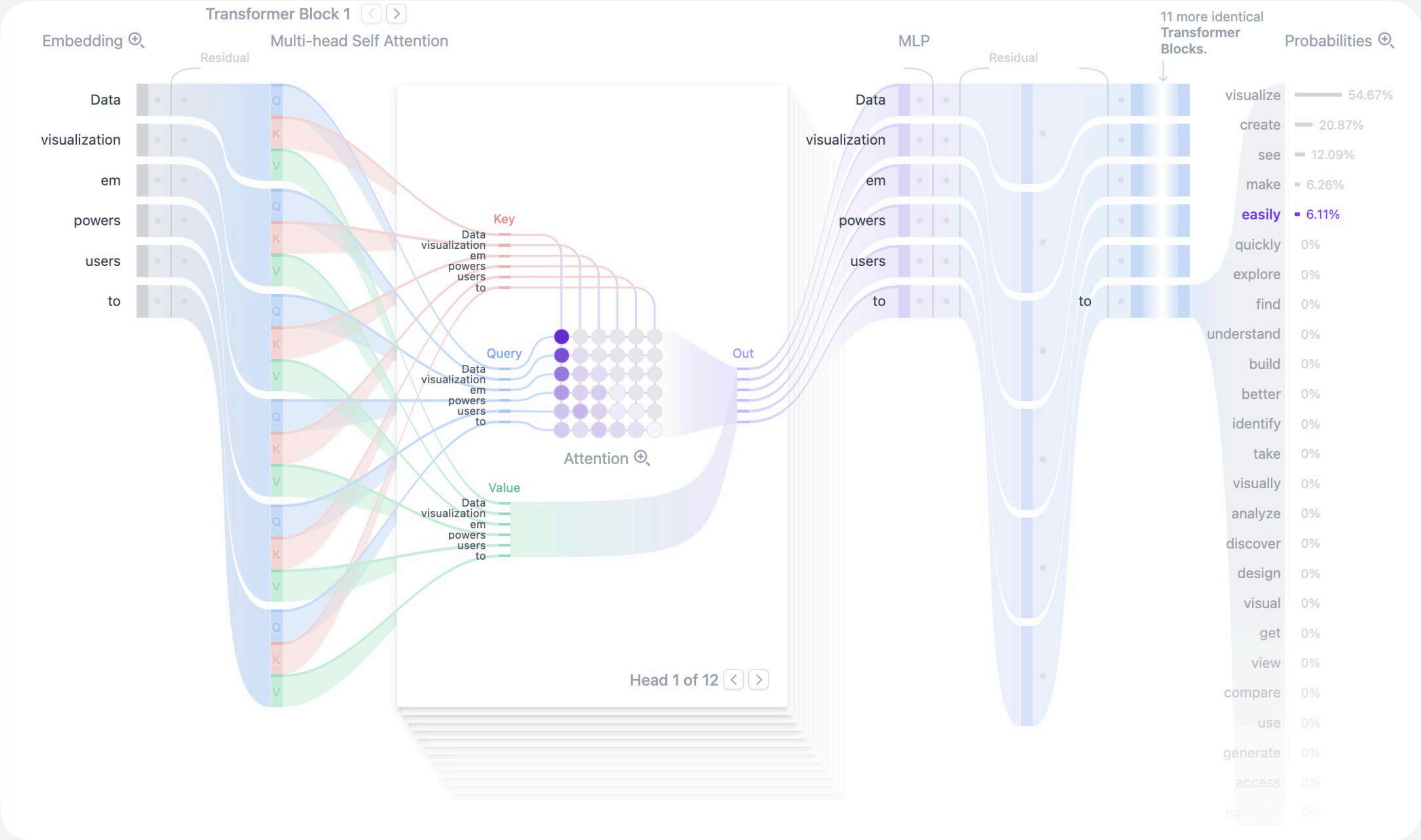
<u>Left</u>: Computational resources allocated to training prominent artificial intelligence systems across various domains, as of March 2025 (Epoch, 2025).

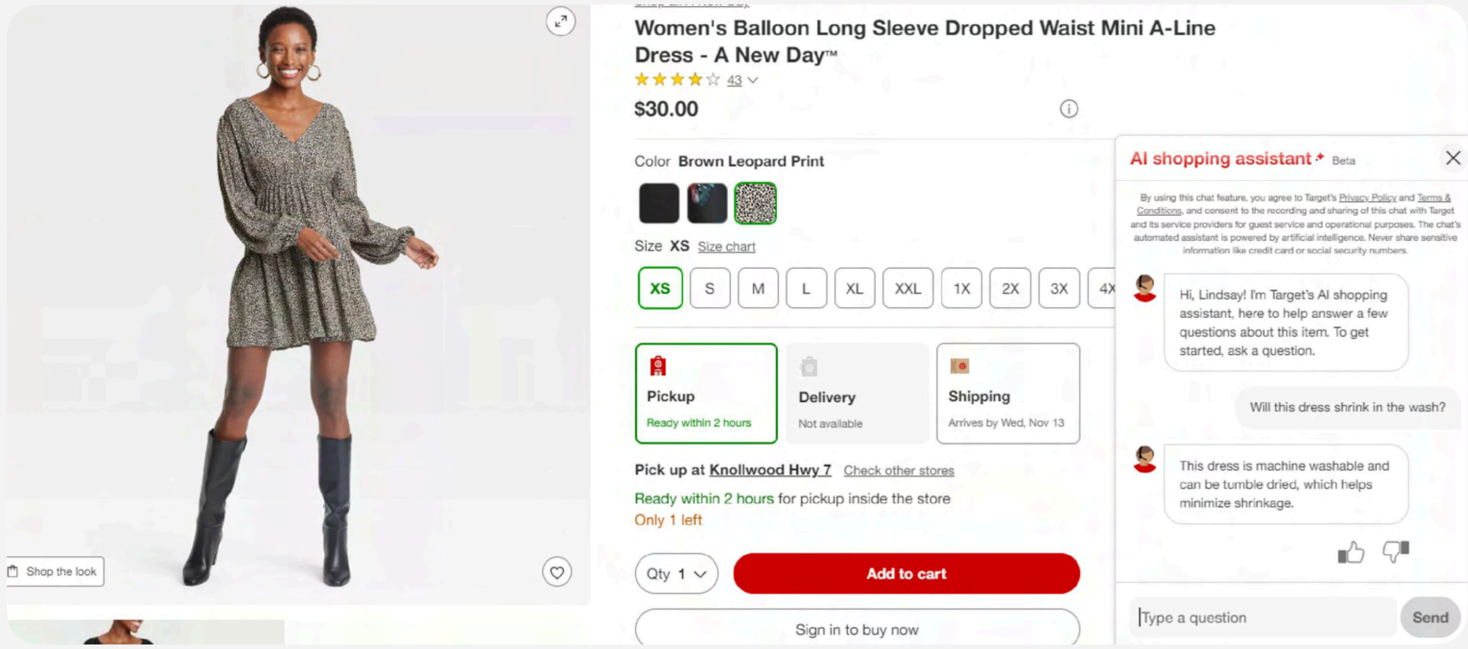
The analysis reveals that since 2021, language model training has consumed the majority of total computational resources. Furthermore, since 2023, language models have emerged as the dominant domain for computational allocation compared to other AI applications.

Living in 2025, you probably have heard enough about AI, a collection of technologies that simultaneously evokes utopian promises and dystopian concerns. Among these innovations, Large Language Models (LLMs) stand out as the technology that has most tangibly integrated AI into our daily work and personal lives. LLMs are also have been put the most computation in training among other artificial intelligence systems. Key industry players of LLMs such as OpenAI articulate missions centered on ensuring that “ensure that AGI benefits all of humanity.” (Pillay, 2025) The promise sounds bright and yet, simplified. Such abstract promises function rhetorically to obscure the complex challenges inherent in the multifaceted relationship between human beings and artificial intelligence systems.

Linguist Emily M. Bender offers a precise articulation of the fundamental disconnect between development LLMs and usage of LLMs: “We’ve learned to make machines that can mindlessly generate text, but we haven’t learned how to stop imagining the mind behind it.” (Weil, 2023) This cognitive dissonance underscores the necessity for developing interaction paradigms that align with human ergonomic constraints, cognitive processing limitations, and values.

This section maps the multifaceted impacts of Large Language Models, analyzing both their beneficial applications and their problematic dimensions. By examining specific interaction paradigms that enhance or diminish user experience, we move beyond abstract promises to identify concrete implementation challenges.





Top: **Visualisation of a transformer bloc k(Cho et al., 2024)**
Bottom: **Target AI shopping assistant**

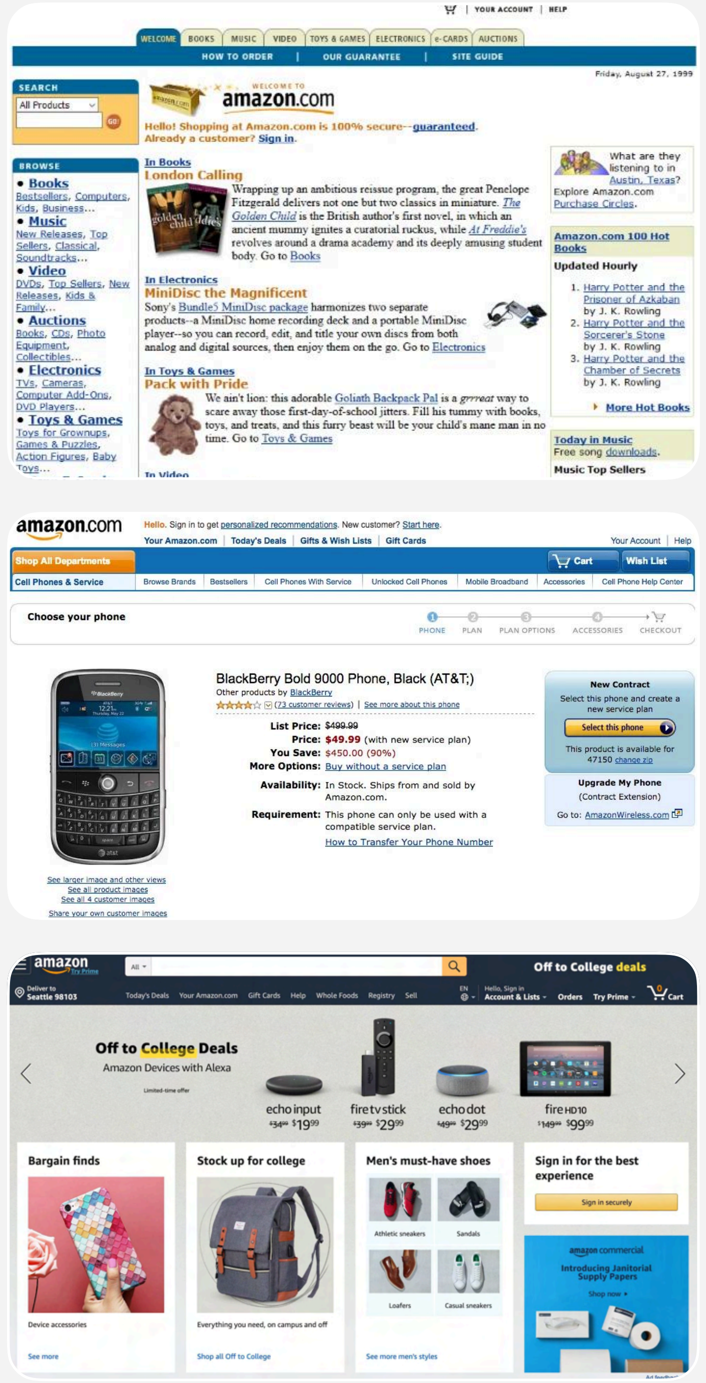
Recent advances in LLM development pivot on innovations in transformer architecture in 2017 (Vaswani et al., 2023), which enables parallel processing of word relationships through self-attention mechanisms. This architectural paradigm has facilitated remarkable scaling capabilities, exemplified by systems such as GPT from OpenAI (Brown et al., 2020). GPT-4 demonstrates remarkable problem-solving capabilities, scoring among the top 10% of test takers on a simulated bar exam (OpenAI et al., 2024). Such performance metrics illustrate the evolution of LLMs from simple text prediction mechanisms to systems capable of complex reasoning tasks—a progression that has catalyzed widespread adoption across diverse industries. In the retail industry, we see new customer-facing applications being introduced. For instance, In 2024, Target released an ”AI Shopping Assistant” on its e-commerce platform . Without having to manually search through user reviews to find such specific product information, shoppers can directly ask a question like ”Will this shirt shrink in the wash?”(Target, 2024). Similarly, True Fit, a company providing solutions for finding right size-and-fit, is evolving beyond its traditional graphical interface to embrace conversation—transitioning from size tables to a generative AI chatbot where shoppers can naturally express their needs with questions like ”What jeans work better for people with muscular thighs?” (Forristal, 2024)

The following section analyzes these use cases and highlights the strategic positioning of LLMs within contemporary e-commerce ecosystems.

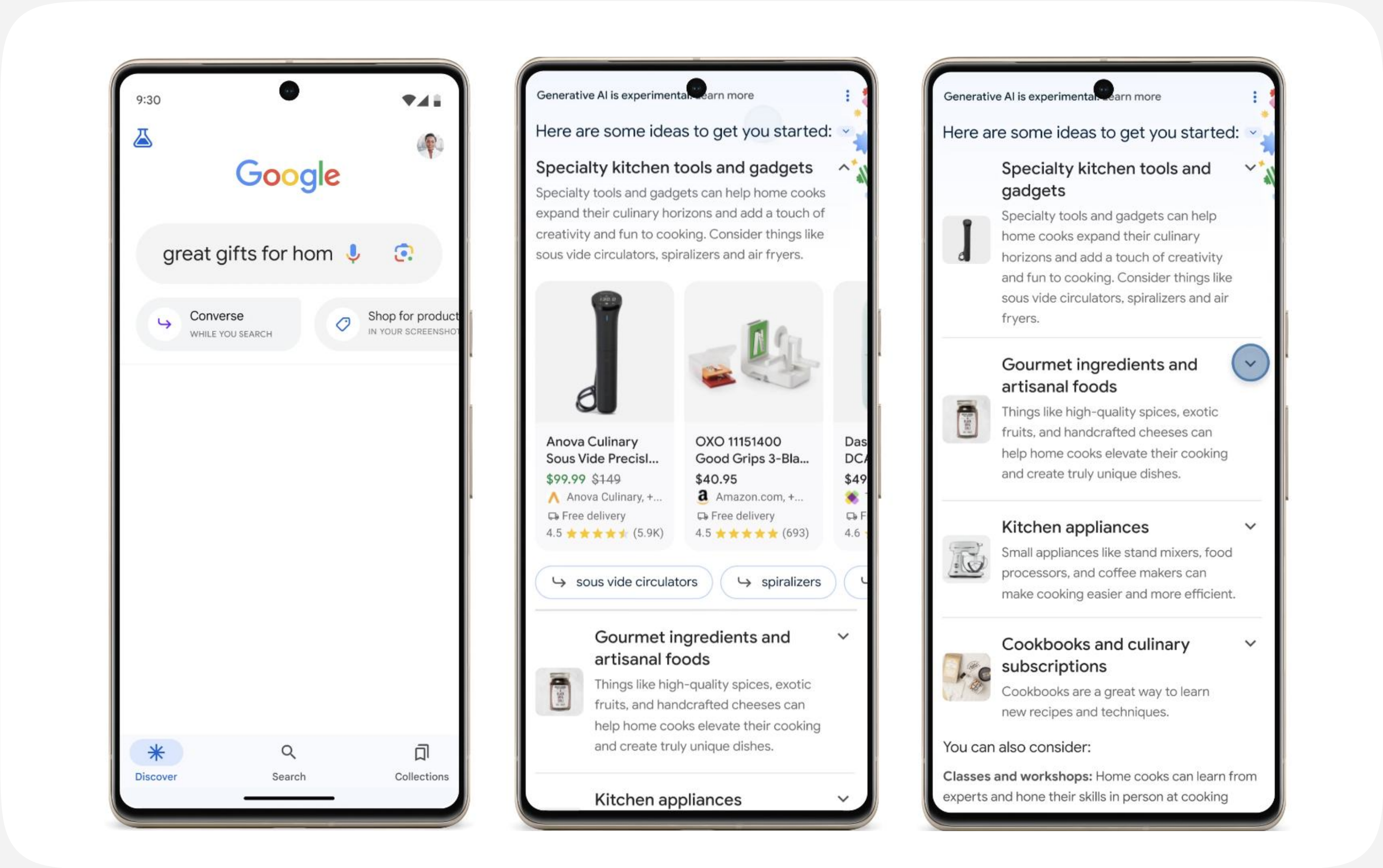
LLM-based systems facilitate dynamic taxonomic alignment between consumer conceptual models and organisational classification schemas, thereby optimizing navigational pathways from expressed user needs to contextually relevant product recommendations.

A significant capability of Large Language Models in the retail context is their capacity to mediate between structured organizational taxonomies employed by retailers and the heterogeneous taxonomic schemas that characterize consumer search and discovery behaviors. A recent study (Cheng et al., 2024) demonstrates how an LLM-based architecture integrating domain-specific fine-tuning techniques with chain-of-thought (CoT) prompting protocols achieves a precision metric of 0.972 in product categorization across granular taxonomic hierarchies. This technical capability enables implementation of user interfaces where imprecise, natural language product descriptions can be algorithmically mapped to structured organizational taxonomies, thereby facilitating more intuitive navigation paradigms within e-commerce systems while maintaining categorical precision.

The capability represents a powerful building block for creating intuitive user experience in e-commerce environments. If we zoom out, we could see that the evolution of e-commerce platform echoes the development of web technologies. Take Amazon as an example, emerging during the Web 1.0 era, the platform initially offered basic HTML interfaces and hyperlink-based navigation reflecting the information-browsing paradigm of early web architecture. The Web 2.0 transition introduced user-generated content mechanisms like product reviews and rating systems. Back to today, the current transition toward Web 3.0 introduces semantic capabilities through algorithmic technologies such as LLM-powered classification systems, potentially reconfiguring fundamental search, discovery, and evaluation processes within digital retail environments.



Bottom: Amazon homepage interface in 1999, 2009, and 2019 (Version Museum, n.d.)



Perhaps, the technological developments like product categorisation represent the collective shift in e-commerce navigation design: transitioning from traditional hierarchical taxonomies requiring explicit user selection through predefined category structures toward predictive systems that algorithmically infer navigational intent from natural language inputs. This transformation redistributes cognitive load from users to systems, with the LLM architecture performing taxonomic disambiguation that previously required manual user navigation through explicit category selection interfaces.

This is an observable industry trajectory rather than speculative prediction. Back in 2023, Google demonstrated this approach through its implementation of LLM-powered search capabilities in consumer-facing product discovery interfaces (Black, 2023).

In the demo, users can input conceptually abstract queries such as 'great gifts for home cooks,' prompting the system to perform semantic decomposition into probable categorical intents like 'gourmet ingredients' or 'kitchen appliances'—search refinements that would traditionally require iterative manual filtering within conventional taxonomic structures (Black, 2023).

Left: Google product search demo

LLM-based systems facilitate automated extraction of key informational elements from unstructured user-generated text, thereby potentially reducing cognitive load.

Consumer decision fatigue remains one of the most significant challenges in e-commerce, with consumers frequently experiencing cognitive overload due to the proliferation of product information, user-generated reviews, and choice alternatives. This information density can transform potentially straightforward purchase decisions into cognitively demanding processes. According to a comprehensive market survey spanning 18,000 respondents across 18 international markets (Bradley et al., 2025), approximately two-thirds of consumers report decision postponement or avoidance when confronting excessive option sets or information volumes, while over 50% express decision anxiety regarding suboptimal purchase outcomes.

This decision paralysis represents a critical opportunity for improved navigation systems that guide customers not just to products, but through the complex evaluation process that follows. Specifically for Decathlon, while it's comprehensive offerings attract customers, many arrive with insufficient knowledge to navigate efficiently toward specific products appropriate for their needs. The breadth of selection—a core strength of Decathlon's business model—simultaneously creates potential for information overload, especially for novice athletes or those exploring unfamiliar sports.

67%

Consumers avoid making decisions when confronted too much information

50%

More than 1/2 consumers express anxiety over decision-making

Data: think with google (Bradley et al., 2025)

soft good quality it is very soft and would recommend only flaw is the threading on the pillow cases are a little loose im sure not all manufactured are like that	4	['soft', 'good quality', 'very soft', 'recommend', 'loose', 'threading', 'pillow cases', 'manufactured', 'flaw']	['soft', 'good quality', 'recommend', 'threading on the pillow cases are a little loose']
nice nice color and stretchyfits a little baggy for slim fit	4	['nice', 'color', 'stretchy', 'fits a little baggy', 'slim fit']	['nice color', 'stretchy', 'fits a little baggy', 'slim fit']
nice shirt for the price looks nice pressed out very neat plenty long to stay tucked in	5	['nice shirt', 'for the price', 'looks nice', 'pressed out', 'very neat', 'plenty long', 'stay tucked in']	['nice', 'shirt', 'price', 'looks nice', 'pressed out', 'neat', 'plenty long', 'stay tucked in']

Left: Sampled results of two fine-tuned LLMs (llama-2-70b-chat and gpt-3.5-turbo-1106) detecting lexical elements from the bodies of user generated reviews.

Large language models demonstrate significant capability in textual summarisation tasks (Lewis et al., 2021), positioning these systems as potentially valuable tools for addressing information overload challenges. Industry implementations increasingly leverage LLM-based systems to distill voluminous product-related content into cognitively manageable, contextually relevant insights.

Recent research demonstrates that fine-tuned LLMs can effectively identify key lexical elements in product reviews that influence customer ratings. For example, when analysing a review stating ”nice hiking shoes it fits well i returned it because i felt the sole wasnt padded enough and the bottom was a bit stiff” (typography preserved from original text from the paper), GPT-3.5 correctly identified the critical phrases ”nice,” ”fits well,” ”sole wasnt padded enough,” and ”bottom was a bit stiff” as the determinants of the 4-star rating (Roumeliotis et al., 2024).

This capability illustrates the potential for deploying LLMs as cognitive offloading mechanisms that systematically redistribute information processing demands from consumers to automated systems during complex decision-making processes. Such implementations could extend user assistance beyond initial product discovery to facilitate comparative evaluation, contextual information augmentation, and cross-validation of alternatives—functions that typically impose substantial cognitive load during the consideration phase of consumer decision journeys across multidimensional choice architectures.

LLM-based systems demonstrate contextual sensitivity, extracting and interpreting situational cues to generate adaptive responses calibrated to specific interaction environments.

Beyond the challenges of navigation and evaluation, personalisation presents a distinct dimension requiring optimisation in e-commerce user experience design. In the context of Decathlon's diverse sporting goods marketplace, the shoppers vary from novice to domain experts. This knowledge heterogeneity influences both the optimal interface strategies for facilitating product selection and the objective product-user fit within specific activity contexts.

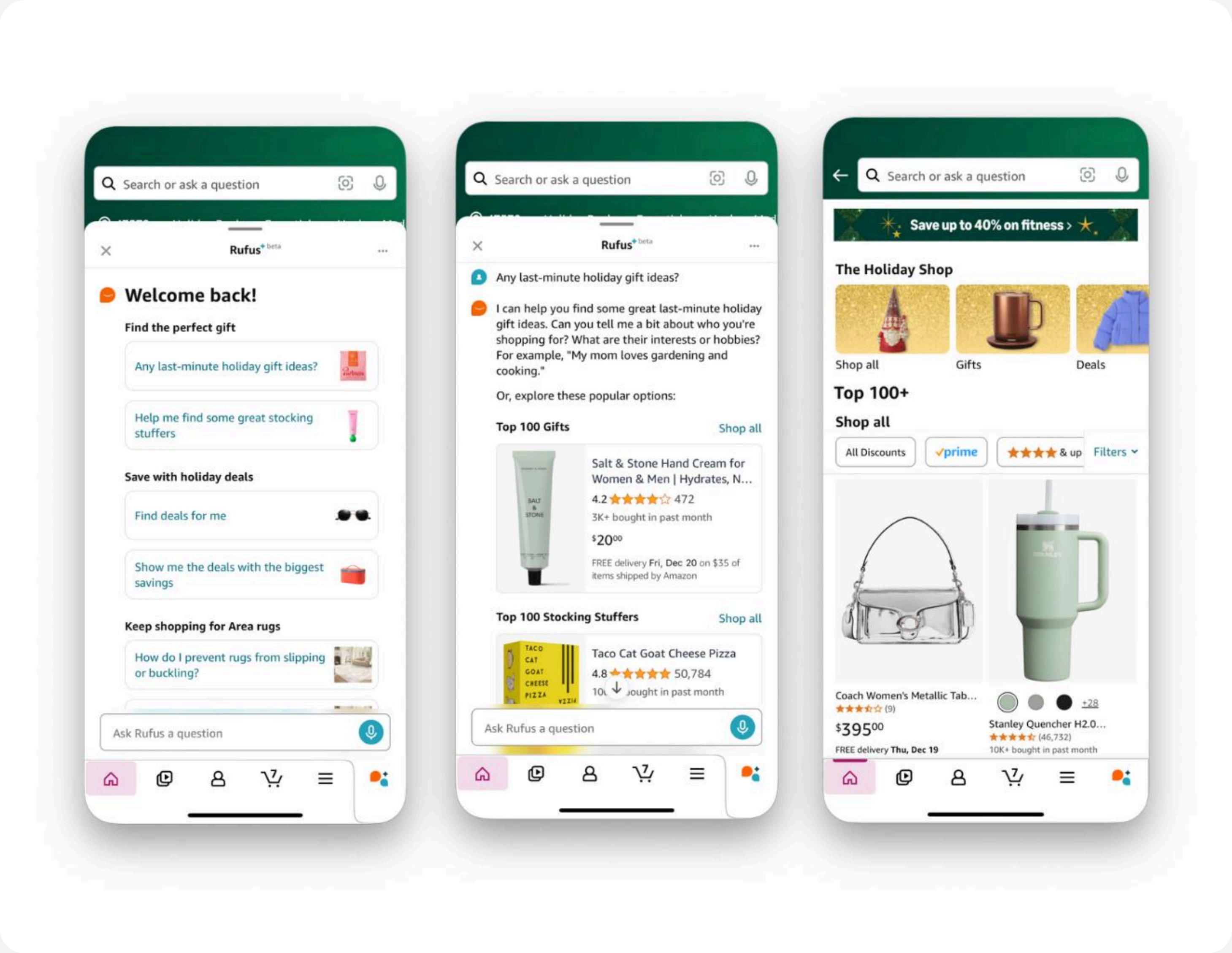
Contemporary research increasingly explores the integration of LLMs with contextual information systems. The term 'context' encompasses multiple definitional dimensions across domains like product design (Hekkert & Van Dijk, 2011)and recommender system engineering (Adomavicius et al., 2022) , but it has been collectively acknowledged as valuable addition to the design of LLM-based systems. More specifically, in the context of e-commerce, LLM-based systems demonstrate promising capabilities for addressing the incorporation of context by adapting generated content when integrated with memory and contextual awareness, especially through Retrieval-Augmented Generation (RAG) (Lewis et al., 2021). A compelling illustration of this approach comes from Google Cloud's Cymbal Shops implementation (Bourgeois, 2024), which demonstrates how RAG enables more contextually relevant product recommendations. Without RAG, the system would process only the description interpreted from user input, whereas with RAG, the prompt incorporates factual data about specific products available in the inventory database. Similarly, Retail-GPT (Arslan & Cruz, 2024) incorporates memory mechanisms that maintain awareness of the user's conversation history and current cart state. When a user in the Retail-GPT demonstration asks about wine recommendations, the system demonstrates its ability to remember both previously discussed products and the current cart contents when formulating its response.

The adaptive architecture of RAG-enhanced systems represents a significant advancement beyond traditional static interfaces in e-commerce environments. While conventional product information pages present standardised content regardless of user characteristics, a more advanced information retrieval architecture like RAG represents the potential of adapting results for specific shopping scenarios, where nuanced consumer characteristics—particularly domain expertise gradients—can be considered to create progressively personalised information presentations.

Concurrently, emerging technical frameworks such as Model Context Protocol (MCP)(Anthropic, n.d.) facilitates and encourages unprecedented interoperability between LLM systems and diverse external platforms spanning development environments (e.g., GitHub), communication tools (e.g., Slack), and geospatial services (e.g., Google Maps). This cross-domain connectivity introduces novel dimensions for personalisation through contextualisation.

While e-commerce-specific implementations leveraging these protocols remain limited in current deployment landscapes, existing LLM-based commercial systems such as Rufus from Amazon demonstrate significant adaptability through dynamic user interface generation workflows (Padurariu, 2025) that integrate both user-articulated inputs and system-derived contextual information. These implementations transcend simple query-response patterns by contextually associating informational responses with relevant product recommendations, creating more sophisticated interaction paradigms that blend informational and commercial functions.

<u>Right</u>: Sampled screenshots of Amazon Rufus interface



What often fascinates me as a designer is how ‘form follows function’ has catalysed countless good design, it inspires me to re-think digital product design.

From music and news to food and clothes, we make choices regularly. How we search for information, understand it, form attitudes, and create memories varies widely between people, influenced by our unique motivations and decision-making approaches (Hoyer et al., 2016, p. 5). These cognitive variations parallel physical anthropometric differences in human populations, creating analogous design challenges. Just as ergonomic chair design addresses physical variability through dimensional optimization for median or extreme user measurements, digital experience design must "mold" around human cognitive behaviors.

In e-commerce contexts, where diverse cognitive processing styles converge within unified systems for product discovery, evaluation, and acquisition, which specific dimensions of cognitive ergonomics warrant prioritisation in design frameworks? This section maps out critical cognitive considerations for LLM-based system design within e-commerce ecosystems, synthesising insights from cognitive science, human-computer interaction, and consumer behaviour research to establish a foundation for the exploration in creating human-centric LLM-based systems. In the next page, there is an overview of the dimensions.

1. Cognitive Load

Excessive or complex information can overwhelm individuals and lead to information overload (Malhotra, 1982). The analysis evaluates both the potential benefits of LLM systems while also addressing potential limitations.

2. Decision-Making

Modern study of consumer decision-making process reveals that consumers utilise their internal mental system to promote useful behaviours that brings positive outcomes (Suomala, 2020).

3. Affordance

This dimension analyses the application of affordance theory as a critical design principle in LLM-based interfaces, emphasising the importance of creating perceptually salient action possibilities that generate appropriate user expectations prior to interaction.

4. Spatial Memory

Analysis on this dimension reflects on the predominant chat-based interaction paradigm in current LLM applications, identifying potential misalignments between conversational design patterns and human memory architecture.

Cognitive Load

Information load has been extensively examined in e-commerce contexts, with early empirical research demonstrating that cognitive processing capacity becomes exceeded when consumers face more than 10 choice alternatives or more than 15 product attributes (Malhotra, 1982).

Recent experimental studies provide a nuanced perspective on this phenomenon, revealing differential user responses to extensive recommendation sets depending on the perceived source. When presented with identical large-scale recommendation sets attributed to either human curators or ChatGPT, participants demonstrated unexpectedly positive evaluations toward AI-attributed recommendations (Kim et al., 2023). While this finding does not conclusively demonstrate cognitive load redistribution, it suggests that reliance on algorithmic recommendation systems may function as a cognitive offloading mechanism, with users' willingness to defer evaluative judgment potentially serving as a compensatory strategy for managing information processing demands. A comparative interface study contrasting traditional menu-based navigation with conversational chatbot interfaces revealed significantly higher reported cognitive load metrics associated with the latter (Nguyen et al., 2022). The researchers hypothesise that this elevated cognitive burden stems from participants' limited familiarity with conversational interfaces for executing traditionally menu-driven tasks. On established interfaces, users could have an experience of sub-consciousness (Forlizzi & Ford, 2000), with the developed memory schemas enabling largely automatised interaction sequences with minimal attentional resources. Conversely, chatbot interactions necessitate multiple conversational turns, each requiring distinct attentional engagement, information processing, and response formulation—cognitive processes that remain in explicit rather than implicit processing channels. This distinction highlights how interaction modality significantly influences cognitive resource allocation irrespective of informational content.

A dominant pattern observed in many of the LLM-mediated applications we have discussed (Target, 2024; Bourgeois, 2024; Arslan & Cruz, 2024) is the implementation of conversational interfaces. As IBM reports, 84% of companies expect to deploy text-based generative AI assistants with customers by 2025—more than doubling from 42% in 2023 (IBM, 2024).

However, empirical cognitive load assessment reveals contradictory findings regarding actual usability. Contrary to intuitive assumptions, multi-turn conversational interactions potentially impose higher cognitive resource demands compared to traditional interface navigation through established design patterns. Users develop cognitive fluency with conventional interfaces—enabling subconscious task execution through automatised interaction sequences—whereas conversational interfaces require sustained attentional engagement across multiple interaction cycles.

Decision-Making

Contemporary research on consumer decision-making illuminates how individuals strategically allocate cognitive resources during information processing in retail environments. Facing often options that risk our cognitive overload, consumers often utilise strategies to simplify the process or the principle of Occam's Razor(Suomala, 2020). In the process, they connect the sensory information in environmental context to their prior belief(Suomala, 2020). In this process, both contextual cues and prior belief function as complementary determinants in decision formation.

For contextual cues, when it comes to products specifically, consumers are sensitive to both intrinsic and extrinsic cues during shopping (Jamal & Goode, 2001). Intrinsic cues relate to physical product attributes—the design aesthetics, material texture, or functional features of a backpack. Extrinsic cues include contextual signals like price points, brand associations, and manufacturing information.

Our reliance on different types of cues varies based on our product expertise. Marketing researchers have revealed how consumers' expertise in specific domains reshapes which cues are prioritized: experts often engage in search activity, but instead leverage their extensive domain knowledge to evaluate intrinsic product attributes more effectively (Solomon et al., 2013). Other consumer characteristics such as personality and demographics also change how we seek and process information cues (Creusen, 2015).

In regard to prior beliefs, often some common market beliefs are utilized for making "mental shortcuts", for instance, some people may believe that the best products are the products that receive the most user reviews, or the higher the price is, the higher the quality of the product is. These simplified decision rules enable efficient navigation of complex choice environments while minimizing cognitive resource expenditure.

In summary, consumer decision-making represents a complex, heterogeneous process influenced by individual differences in cognitive processing styles and domain knowledge. Despite this variability, the principle of cognitive simplification emerges as a unifying design consideration: systems can optimize information processing by facilitating connections between presented attributes and users' existing belief structures. This alignment reduces cognitive load by leveraging rather than contradicting established mental models. For instance, novice consumers who employ the review quantity heuristic as a quality proxy would benefit from prominent presentation of review volume metrics, whereas expert users might require more substantive quality indicators aligned with their more sophisticated evaluation frameworks.

Affordance

Another fundamental misalignment between human cognitive architecture and chat-based interfaces involves mental model formation. This concept has deep theoretical roots in cognitive psychology and human-computer interaction, notably articulated in Norman's seminal work on design psychology (Norman, 1988).

Norman established that interfaces generate implicit functional expectations through their affordances—a principle particularly relevant to conversational systems. Chat-based interfaces, through their anthropomorphic presentation and conversational interaction patterns, inadvertently establish expectations of human-equivalent comprehension capabilities, or that they are monitored by real human beings. These expectations create a cognitive discontinuity when users encounter the actual limitations of the underlying statistical language processing systems.

Empirical research (Luger & Sellen, 2016) demonstrates that when interacting with chat-based systems, users form these unrealistic expectations primarily through interface cues rather than actual system capabilities. Their interviews with conversational agent users revealed that participants, especially those lacking technical knowledge, consistently expected human-like intelligence based solely on the presentation of the interface—in the way of human-like sentences and the tendency toward anthropomorphism. When these expectations were violated, users experienced significantly higher frustration compared to interfaces that established more accurate capability models. The study by Hohenstein and Jung (2020) demonstrates that users attribute not only agency but moral responsibility to conversational agents based on interface presentation. This shows how deeply conversational interfaces shape how we mentally model what systems can do.

This misalignment becomes particularly problematic in e-commerce contexts where product selection involves complex multidimensional decision-making.

These limitations, though temporarily masked when the experience exceeds user expectations, could possibly damage users' trust on the brand and user satisfaction when interactions inevitably fall short. Thus, it is essential to explore alternative interaction paradigms that incorporate conversational intelligence while providing users with a more accurate mental model of system capabilities and limitations.

Spatial Memory

A final critical misalignment between chat-based interfaces and human cognitive architecture involves information organization paradigms and their compatibility with spatial reasoning processes. Foundational cognitive research on spatial mental models (Tversky, 1993) established that human cognition relies extensively on spatially-organized representational structures for comparative evaluation, relational analysis, and information retention—precisely the cognitive operations that dominate complex shopping decisions requiring multi-attribute comparison across product alternatives.

Traditional e-commerce interfaces leverage this spatial cognition effectively. Throughout the shopping journey: product grids allow simultaneous comparison across options, specification tables enable direct feature comparison, persistent filtering controls maintain visibility of decision parameters, some e-commerce platform adopts design patterns like previously viewed items, which also provides the spatial arrangements that directly connect users' mental model of relevancy to the interfaces. These interfaces align with how humans naturally organise information for complex decision-making, provide critical decision scaffolding, facilitate the creation what Norman terms a more accurate ”conceptual model” that matches user cognitive processes (Norman, 1988).

The predominantly linear, temporally-sequenced information presentation in conversational interfaces fundamentally conflicts with these spatial cognitive strategies, requiring users to maintain multiple information elements in working memory rather than offloading this cognitive burden through spatial distribution of information. This forces users to maintain product information, their thought process in the user journey with chat-interface entirely through working memory, creating substantial cognitive load.

Emerging interface patterns in generative AI tools provide valuable insights for addressing these limitations. Tools such as canvas in ChatGPT (OpenAI, 2024) and artifacts in Claude (Anthropic, 2025), introduce spatial organization to AI interactions by allowing multiple conversation threads to exist simultaneously in a persistent visual space.

These interfaces enable users to arrange information spatially, preserving context across interactions and supporting the natural comparative processes humans employ during complex decision-making. Such approaches suggest promising directions for hybrid e-commerce interfaces that could combine conversational accessibility with the cognitive advantages of spatial organisation, enabling users to more easily track information sources and visualise their progression through the shopping journey.

The challenges and opportunities co-exist.

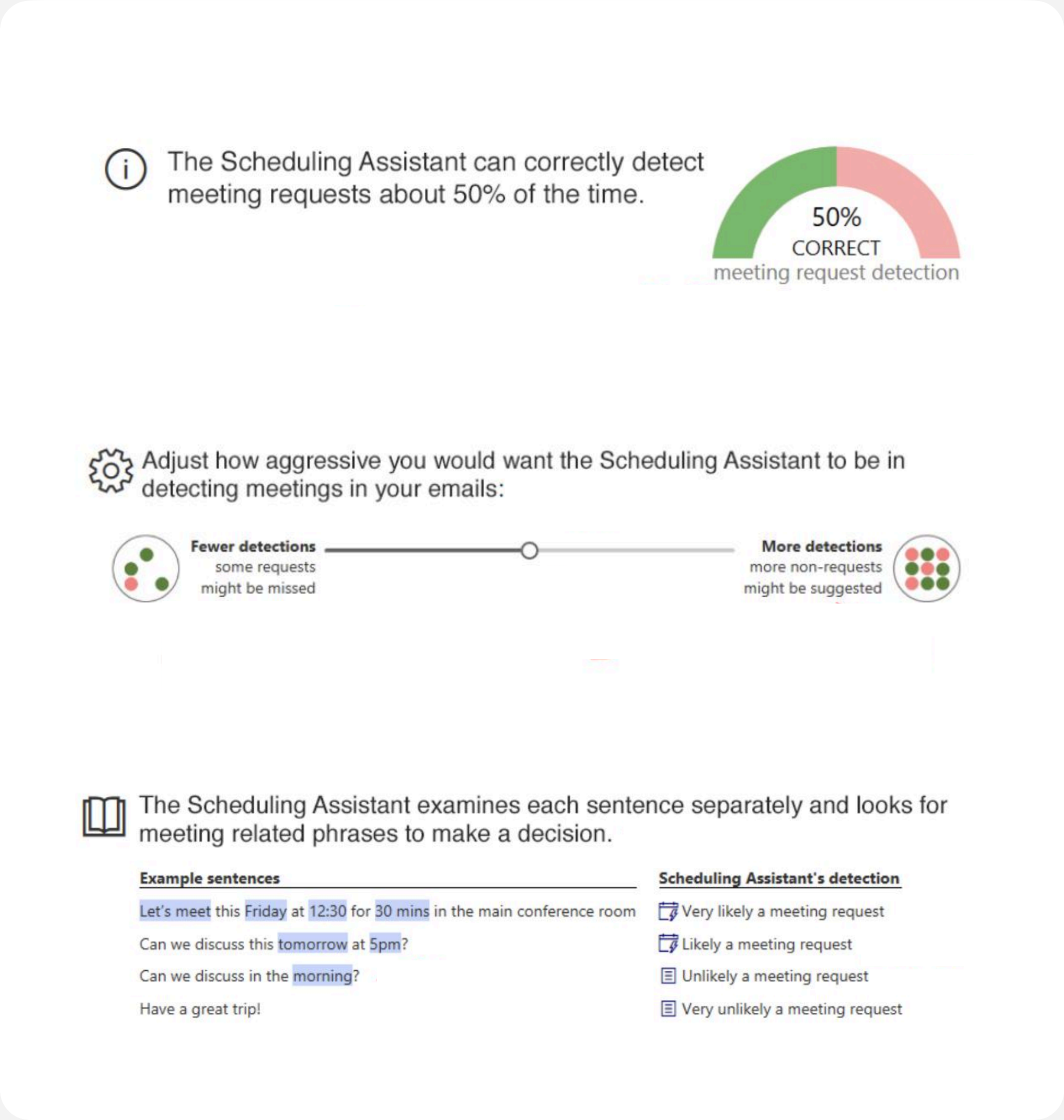
To conclude, current commercial implementations demonstrate promising capabilities in supporting certain consumer cognitive processes-LLMs show particular strength in dynamic categorisation alignment and information summarisation, addressing some aspects of consumer memory limitations and expertise differences. However, these implementations, predominantly relying on conversational interfaces, only partially address the full spectrum of cognitive processes involved in shopping decisions. While they excel at natural language understanding and personalisation, they often struggle with supporting spatial comparison strategies, managing working memory limitations during complex decisions, and providing appropriate affordances that align with user expectations.

LLMs enabled new ways of searching, navigating through purchase options within digital commerce environments. The technology advantages might inspire designers to create elegant interface solutions that provide users with flourishing personalised experience, but also challenge designers to overcome the increased cognitive load and inaccurate conceptual models that misalign with the system capabilities.

Existing approaches aiming for overcoming these challenges.

Right: Design patterns from the paper "Will You Accept an Imperfect AI?"

Emerging empirical studies have approached AI interface re-design and evaluation on the foundation of established human-computer interaction theoretical frameworks, emphasizing the importance of generating appropriate user expectations through interface affordances. One promising approach for addressing this gap comes from a study in 2019 "Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems" (Kocielnik et al., 2019). The researchers from University of Washington and Microsoft designed alternative interfaces and evaluated them through methods that are grounded in the Expectation Confirmation Model (ECM). The ECM model sets the premise that the expectations users have towards the system directly affects the satisfaction and acceptance of the actual experience with the system. They designed and tested multiple expectation-setting techniques with a scheduling assistant, finding that three design patterns—explicit "accuracy indicators" that directly communicate expected system accuracy, "example-based explanations" that help users understand how the system works, and "user controls" that give users agency over system behavior—significantly improved user satisfaction and acceptance of the designed AI systems. Their experimental comparison confirmed these design components created lower expectation discrepancy and higher perception of understanding—ultimately creating more appropriate end-user expectations.



In another study that aims to facilitate better understanding of AI systems for the end-users (Springer & Whittaker, 2019), the researchers implemented an interactive explainability interface that decomposed AI predictions into components that users could independently probe for more fine-grained transparency. However, contrary to design expectations, empirical evaluation revealed that this granular decomposition approach actually impaired user experience by creating attentional fragmentation. Based on these findings, the researchers recommend exploring simplified heuristic transparency mechanisms that provide appropriate system visibility without violating user expectations or introducing excessive cognitive overhead during primary task execution.

Rather than accepting these limitations as inherent to LLM integration, we propose to follow the approach demonstrated in Kocielnik et al.'s work (Kocielnik et al., 2019) and Springer and Whittaker's work (Springer & Whittaker, 2019) by designing new interface paradigms specifically engineered to bridge these gaps. Drawing inspiration from these empirical investigations, the project was reframed into LLM interface design exploration that would guarantee proper expectation management as a central principle, which is supported by the previously discussed user-centric design considerations such as cognitive load, decision-making, affordance, and spatial memory, creating interaction patterns. This approach enables evaluation through established metrics that quantify expectation-experience gaps, measuring discrepancies between users' anticipated system behavior and their actual interaction experience.

Within this analytical framework, three primary research questions emerge:

1. *How might we leverage LLM technology to improve* navigation experience *for decathlon website users?*

2. *How do non-conversational LLM interaction patterns* affect user experience *compared to purely conversational interfaces?*

3. *How do non-conversational LLM interaction patterns* manage user expectations *compared to purely conversational interfaces?*

The initial research question establishes a human-centered design inquiry focused on applying LLM technologies within digital commerce environments while respecting cognitive ergonomic principles. The subsequent evaluation questions examine design outcomes through complementary analytical lenses: the second question addresses functional user experience dimensions, while the third specifically evaluates expectation management efficacy—directly applying the theoretical foundation constructed around expectation-experience alignment as a critical evaluation metric.

Conceptual map of the related work.

Create Meaningful Expectations

Minimise the gap between the expectation and perceived experience towards LLM-based systems

Potential of LLMs

1. Taxonomy alignment
2. Information Extraction
3. Contextual sensitivity

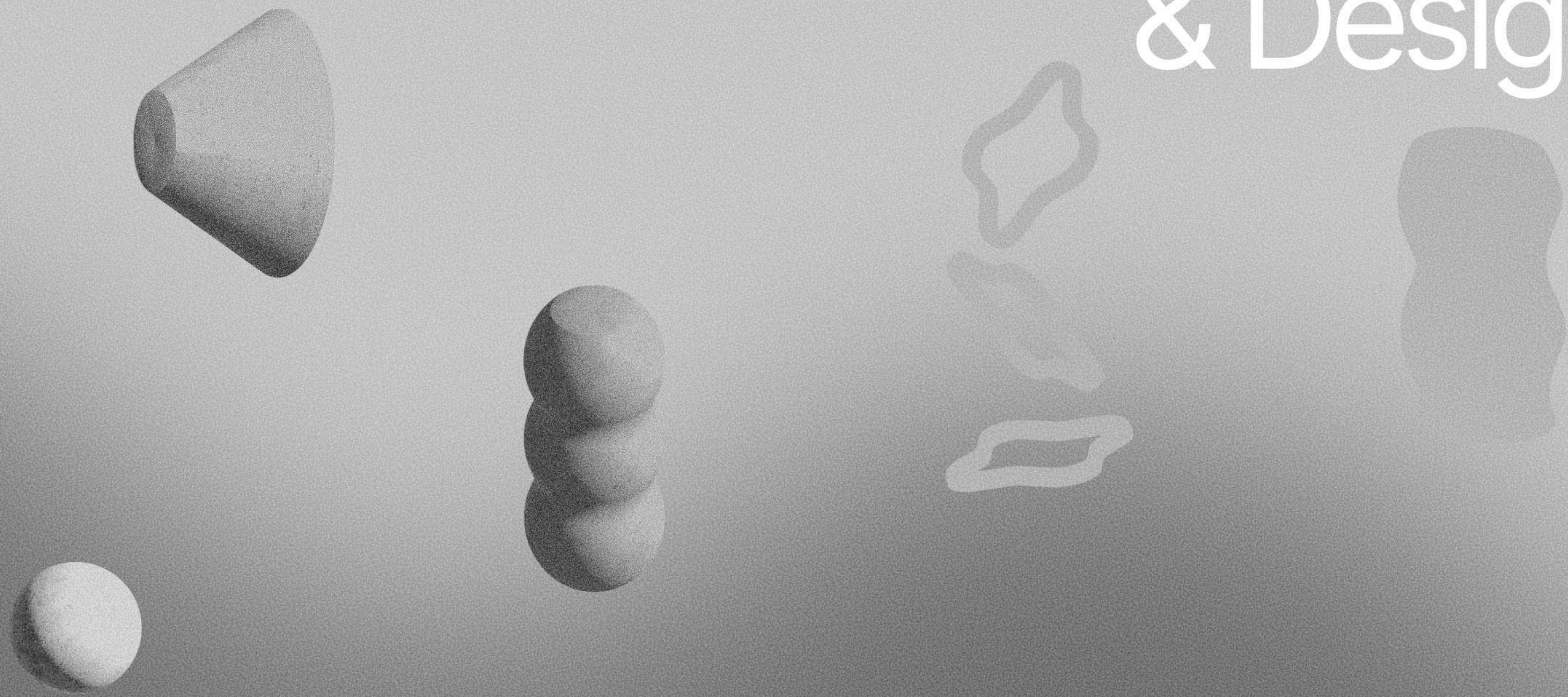
Should transcend current limitations

Addresses design chellenges

Limitations of Chat-based Interfaces

1. Cognitive Load
2. Affordance
3. Decision-Making
4. Spatial Memory

Discover & Design



3.1

Design Approach

The methodological approach adaptations to work with conversational AI as a design material

3.2

Context Understanding

Ground the design objectives into the context of touch points of Decathlon

3.3

Early Experiments

A few experiments that helped me with grasping the dimensions in designing and examine user experience with LLM

3.4

Design Iterations

Iterative prototyping that is focused on improving usability details

In the previous chapter, we had the discussion around the capabilities and limitations of conversational AI, with the challenges in mind, rather than using AI capabilities as the starting point for innovation, I chose to explore how to design something inherently intuitive for users. This motivation led me to adopt User-Centered Design (UCD) as my overall design approach —a methodology that places users' experiences at the foundation of creation. UCD delves into understanding what potential users truly need, aspire to, and are capable of when interacting with a product, service, or system (Van Boeijen et al., 2020).

“Design thinking is ...integrate the needs of people, the possibilities of technology, and the requirements for business success.”

Tim Brown

Executive Chair
of IDEO

“Let us get used to looking at the world through the eyes of others.”

Bruno Munari

Designer

Adapting UCD for LLM-as-Material

The UCD approach follows five key steps: (Van Boeijen et al., 2020)

- Front-end User Research - Through methods like user observations, context mapping, and interviews.
- Define - Converge findings from the front-end user research into a defined scope through problem definitions and personas.
- Design - A creative process that integrating suitable methods from storyboarding to user journey mapping.
- Prototype - Externalisation ideas through prototyping.
- Evaluate Use - Evaluate the design outcomes via usability testing to assess how the product fulfils users' needs and provides value.

What sets this project apart from my previous projects is treating the language models not just as tools, but as core design materials - just like how we'd work with typography, color, or form. Instead of only designing with traditional elements like shapes and arrows, the language models became part of my design palette. This shift meant I had to adapt my entire UCD process to account for these 'living' materials that behave differently than static design elements.

Since I was designing with materials that actively respond to user input, my user research became laser-focused on understanding those interaction patterns. My research extended beyond typical interaction patterns to include some surface-level linguistic analysis. Without established user behaviours to reference - like you'd have with physical products - I had to identify how people naturally formulate questions and communicate their underlying intentions in the interaction with the commerce system.

The creative and prototyping phases demanded a completely different approach. Since the LLM generated content became both my research material and a functional component of the user experience, I had to learn to 'sculpt' with language models - experimenting with prompts and responses the way I'd normally explore texture, form, or color.

Ultimately, evaluation became a delicate balance between exploring the untapped potential of LLM-based interfaces and maintaining research rigor. Over-constraining user interactions would limit our understanding of these emerging systems' capabilities and how we could better shape them, but without structure, systematic comparison and meaningful insights became impossible.

Value Mapping

Despite that UCD methods have been invaluable in guiding my creative process of designing with LLMs and understanding how to solve actual user problems, my perspective is biased towards that technology is not always positioned as inherently valuable for users—rather, we can use it to elevate user experiences in meaningful ways. o better understand these dynamics from both the researcher and business stakeholders, I created a value mapping framework that illustrates where user needs, company objectives, and my research goals align and diverge. This mapping helps identify the ideal design space where all stakeholders' values can be honoured while maintaining a user-centred approach.

AI INNOVATION PERSPECTIVE

"To explore human-AI interaction design principles "in the wild" (Amershi et al., 2019)"

"To incorporate societal context and values into AI systems (Rahwan, 2017)"

SOCIOLOGICAL PERSPECTIVE

"To explore alternative to the focus on humanoid embodied intelligence (Cave et al., 2018)"

BUSINESS PERSPECTIVE

"Recreating customer experience (Decathlon, 2024)"

"Minimise changes on interface architecture (Past Interview)"

"User interactions are part of co-created brand image, e.g., through word-of-mouth (Beverland, 2018)"

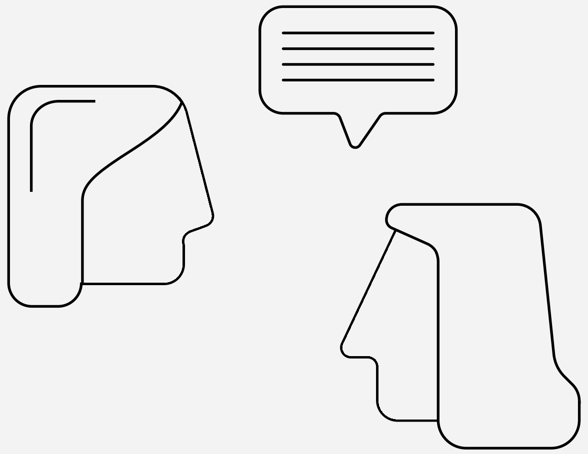
"AI should show helpfulness and generosity to customers (Stakeholder Interview)"

UCD PERSPECTIVE

"Innovation using user-centric design approach (Van Boeijen et al., 2020)"

"Technology should be intuitive and augment humans (Norman, 1988)"

What can we learn directly and indirectly from end-users and stakeholders?



My initial user research combined two complementary approaches to capture both authentic behavior and expert insight. First, I conducted immersive observations at Decathlon's Bijmer Arena, positioning myself as an in-store assistant to document genuine customer questions and staff interactions within the actual shopping environment.

Second, I interviewed a seasoned Decathlon professional with extensive product and customer knowledge, gathering their assessment of various LLM-generated responses to understand what constitutes helpful assistance from an expert perspective.

Combining direct customer observation with expert response analysis gave me insights into both sides of the human-LLM interaction: how users naturally formulate questions (input) and how different response styles impact their experience (output). This dual perspective provided a foundation for identifying design challenges and opportunities in the LLM space.

When I initially began investigating the problem, I lacked clarity on the choice of research methods. Rather than continuing to diverge broadly, I should have incorporated convergent approaches to better narrow and define the core problem - methods like co-creation sessions and decision matrices. This focused approach would have enabled me to deploy more generative methodologies for steering deeper, more targeted conversations about LLMs. Instead, my findings reflect primarily individual synthesis rather than collective insights, limiting their broader applicability and stakeholder validation.

Field Research: Customer Questions

Capturing natural inquiry patterns in the actual retail environment

To design a solution that genuinely guides consumers through product discovery, I needed to understand the natural questions that reveal their underlying needs and the cognitive processes behind their shopping behavior. I spent half a day immersed in Decathlon's Bijmer Arena, strategically positioning myself as an in-store assistant to observe authentic customer interactions within their natural context. I documented these exchanges between customers and me or the store.

During scheduled breaks, I retreated to document these rich interactions through detailed ethnographic notes, carefully preserving the authentic language customers used to describe their needs. This approach respected privacy considerations by avoiding any audio or video recordings, while still capturing the nuanced ways shoppers navigate the gap between their personal product categorization systems and the store's organizational logic.

REFLECTION POINT

These field observations revealed that customers don't just ask for products - they're constantly navigating cognitive gaps between their personal categorization systems and the store's organizational logic. When faced with this mismatch, shoppers naturally develop bridging strategies: they describe products using their own terminology, reference specific use cases, or ask for guidance to unfamiliar sections. This adaptive behavior suggests that effective LLM-based assistance must go beyond simple product matching to actively help users translate between consumers' mental models and the retail structures.

"Which shoes are waterproof?"

While I was working in the hiking department, a young male customer approached me with a direct question about which shoes were waterproof. Not being immediately certain, I consulted my colleague who informed us that '90% of these shoes are waterproof.' This interaction illustrates how some customers prefer to efficiently narrow their options through targeted questions rather than individually examining each product's specifications.

"Are the helmets all here? or you have separate section for kids' skiing?"

While I was near the winter sports department, a female customer shopping with her family approached me asking this question. This question reveals several possibilities: her thoughtfulness in seeking age-appropriate options, difficulty finding suitable children's helmets in the main display, or prior experience with stores that separate adult and children's merchandise. Regardless of her specific motivation, this interaction demonstrates how shoppers mentally organise products into categories like 'kids' helmets' and may prefer direct guidance to these specialised sections when available to better compare between them.

"I want to find those long pants for winter"

While I was near the exit, a male customer approached me asking this question. He attempted to clarify by using hand gestures to describe the item. This interaction was notable because his query was quite vague, and the fact that he sought assistance near the end of the aisle suggests he had already searched unsuccessfully and become somewhat disoriented. My colleague directed him to some departments where he might find suitable products. This incident illustrates how consumers without specific knowledge of product terminologies, design features, or materials may struggle to navigate the store environment independently.

"Where is golf?"

This question emerged as one of the most frequently asked according to my colleagues who have worked in-store for more than half a year. During my observation, I was fortunate to be approached by customers with this exact inquiry. The reason behind this persistent question reveals an interesting mismatch in spatial organization: unlike more intuitive product categorizations (where hiking shoes belong in hiking sections and yoga mats in fitness areas), golf equipment is tucked away in a corner of the "other team sports" section alongside basketball and paddle ball items. This placement creates a cognitive disconnect for shoppers whose mental model of store organization differs from the actual layout they encounter.

"Where are the skates?"

Similar to the golf case, a young male customer approached me asking for skate equipment, despite the fact that these items were positioned right next to us at that moment. This highlights how consumers can become visually overwhelmed in product-dense environments. When faced with an abundance of merchandise, shoppers can literally fail to see what's directly in their line of sight—a fascinating example of how sensory overload in retail spaces can create a form of temporary "blindness" to products that are actually within reach

"Where can I find this pedal ball?"

A young female customer approached me, phone in hand with a screenshot displayed, asking if I could help locate the item—a yellow paddle ball. This interaction wasn't unique; multiple shoppers sought assistance finding products they had discovered online.This observation illuminates how many consumers engage in extensive research before visiting physical stores, or rely on past experiences to inform their current shopping journey.

From these questions, we can observe consumer cognitive processes unfolding in real-life shopping environments, revealing the complex mental work involved in navigating retail spaces. When what consumers encounter doesn't align with their personal categorisation strategies, they don't simply abandon their search - instead, they actively develop bridging mechanisms to reconcile this cognitive dissonance. These observations reveal that effective retail assistance - whether human or LLM-powered - must function as a cognitive translator, capable of interpreting diverse mental models and guiding customers through the gap between their intuitive organizational systems and commercial retail logic.

Interview: Expert Perspective

Expert evaluation of LLM response variations

Having examined one side of the LLM system—how users naturally formulate questions—I moved to the other side, the out put, where I needed to investigate the dimension of the generated responses.

What is a good response? Can we detect the nuances between the outputs from different LLMs? Do these nuances meaningfully impact their experience? To explore this, I collaborated with a seasoned Decathlon expert who possesses deep knowledge of the company's product ecosystem and customer behaviours.

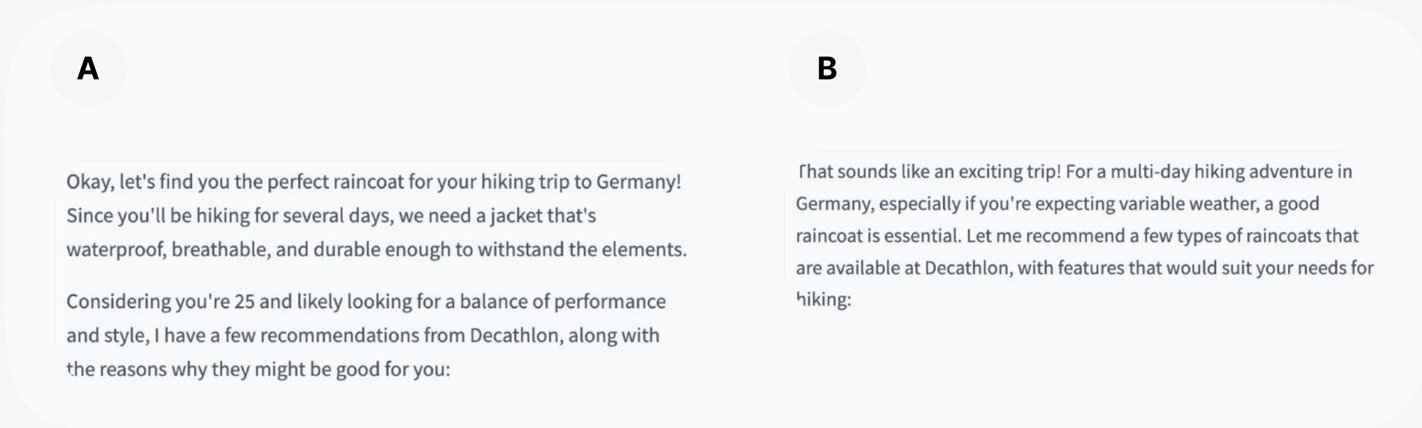
Together, we analyzed responses from two different LLM using a controlled comparison approach. I leveraged Chatbot Arena's side-by-side functionality to examine how different large language models handled the same task: acting as Decathlon shopping assistants recommending hiking raincoats for a specific customer scenario—"a 25-year-old planning a multi-day hiking trip with friends in Germany".

This methodology allowed us to identify the subtle but potentially significant differences in tone, product knowledge, and recommendation strategies that might influence user trust and engagement with LLM-based shopping assistance.

REFLECTION POINT

The expert's assessment validated a fundamental challenge we have discussed earlier, that different customer segments require different information architectures. Novice shoppers benefit from contextual, narrative-driven responses that build confidence, while experienced customers prefer direct access to technical specifications. This suggests that effective LLM-based assistants must not only generate appropriate content but also adapt their communication structure based on inferred user context.

> They both miss something in the beginning, we usually ask clients many questions, like what routes you will take, who you will go with.... AI should be able to ask these questions.


A

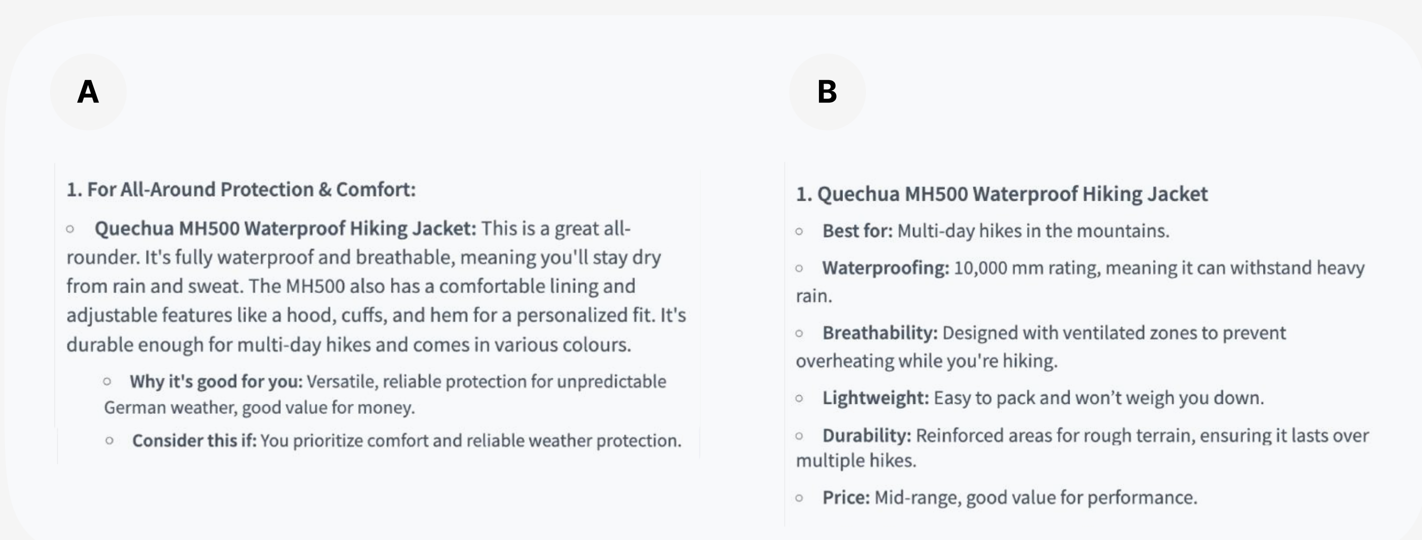
Okay, let's find you the perfect raincoat for your hiking trip to Germany! Since you'll be hiking for several days, we need a jacket that's waterproof, breathable, and durable enough to withstand the elements.

Considering you're 25 and likely looking for a balance of performance and style, I have a few recommendations from Decathlon, along with the reasons why they might be good for you:

B

That sounds like an exciting trip! For a multi-day hiking adventure in Germany, especially if you're expecting variable weather, a good raincoat is essential. Let me recommend a few types of raincoats that are available at Decathlon, with features that would suit your needs for hiking:

> A sounds better in the tone, but it also depends on who you are recommending to, for people who are experienced they want to know the technical details like in B.


A

1. For All-Around Protection & Comfort:

- **Quechua MH500 Waterproof Hiking Jacket:** This is a great all-rounder. It's fully waterproof and breathable, meaning you'll stay dry from rain and sweat. The MH500 also has a comfortable lining and adjustable features like a hood, cuffs, and hem for a personalized fit. It's durable enough for multi-day hikes and comes in various colours.
    - **Why it's good for you:** Versatile, reliable protection for unpredictable German weather, good value for money.
    - **Consider this if:** You prioritize comfort and reliable weather protection.

B

1. Quechua MH500 Waterproof Hiking Jacket

- **Best for:** Multi-day hikes in the mountains.
- **Waterproofing:** 10,000 mm rating, meaning it can withstand heavy rain.
- **Breathability:** Designed with ventilated zones to prevent overheating while you're hiking.
- **Lightweight:** Easy to pack and won't weigh you down.
- **Durability:** Reinforced areas for rough terrain, ensuring it lasts over multiple hikes.
- **Price:** Mid-range, good value for performance.

When comparing the two responses, the expert immediately identified distinct approaches: Option A offered descriptive, contextual guidance while Option B focused on technical specifications. This raised a natural question - which approach would work better in a real AI assistant scenario?

The expert's assessment revealed that effectiveness depends entirely on user expertise. Experienced customers value direct access to technical details like 'breathability ratings' and 'lightweight construction' (Option B), while novice shoppers benefit more from contextual explanations that build understanding and confidence (Option A).

This finding highlights a critical challenge that no single format appeals universally for the users. Different consumers bring distinct information-seeking preferences and expertise levels to their shopping journey - a insight that aligns with patterns identified in the related work research.

Problem Definition

Based on the research insights, there are several key challenges in the retail information-seeking experience:

- **Reducing Information Overload**

  Despite physical product visibility, consumers experience "choice blindness" in product-dense environments. An effective assistant must combat cognitive overwhelm by intelligently filtering and prioritising information relevant to specific shopping contexts.

- **Connecting Shopping Journeys**

  Customers frequently arrive with prior online research - showing product screenshots or referencing items discovered digitally. This requires seamless continuity between digital discovery and physical navigation, helping users locate and compare previously researched items.

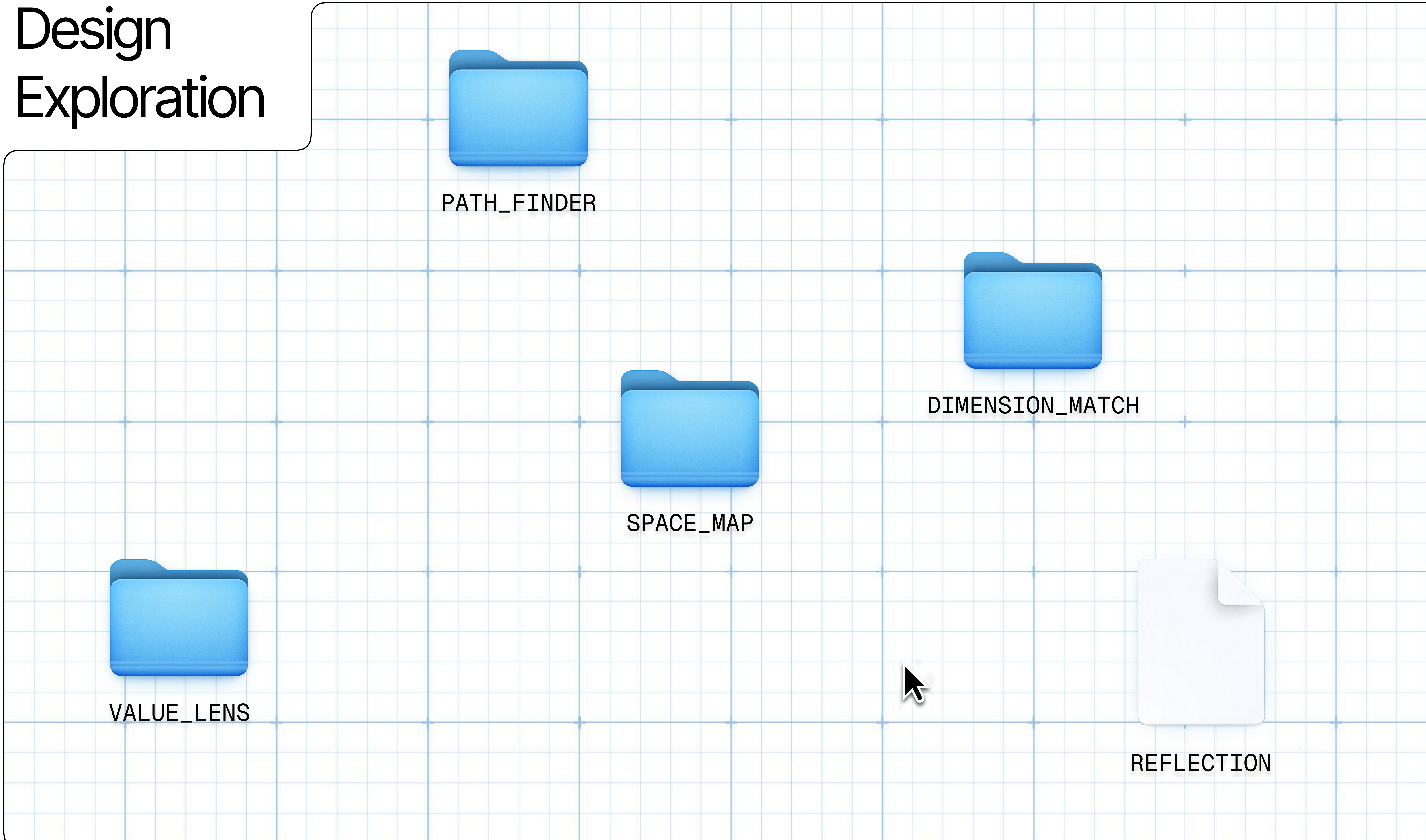
- **Interpreting Diverse Language Patterns**

  Users struggle when expressing needs through diverse linguistic patterns (from specific queries like "Which shoes are waterproof?" to vague descriptions like "those long pants for winter") and when store layouts don't match their mental organisation systems (as seen with "Where is golf?" questions). This necessitates an assistant that can both understand varied terminology and translate between personal classification systems and official store categorisations—leveraging LLMs' ability to extract intent from ambiguous queries and understand semantic relationships across different taxonomies.

- **Adapting to Expertise Diversity**

  The expert evaluation revealed a critical divide: novice users need descriptive, confidence-building explanations while experienced customers prefer direct technical specifications. Effective AI assistance must detect user knowledge levels from conversation patterns and dynamically adjust information density accordingly.

Design Exploration



Development of four early-on prototypes

With these problems in mind, I began designing quick prototypes through creative exploration. In the following pages, I detail a series of experimental prototypes exploring how LLMs can address the identified challenges. To make the prototypes easier to remember, I gave them descriptive names:

PathFinder assists with personalized product guidance, ValueLens uncovers users' deeper motivations through structured dialogue, SpaceMap provides spatial orientation through interactive visualization, and DimensionMatch evaluates items along customized dimensions relevant to individual preferences.

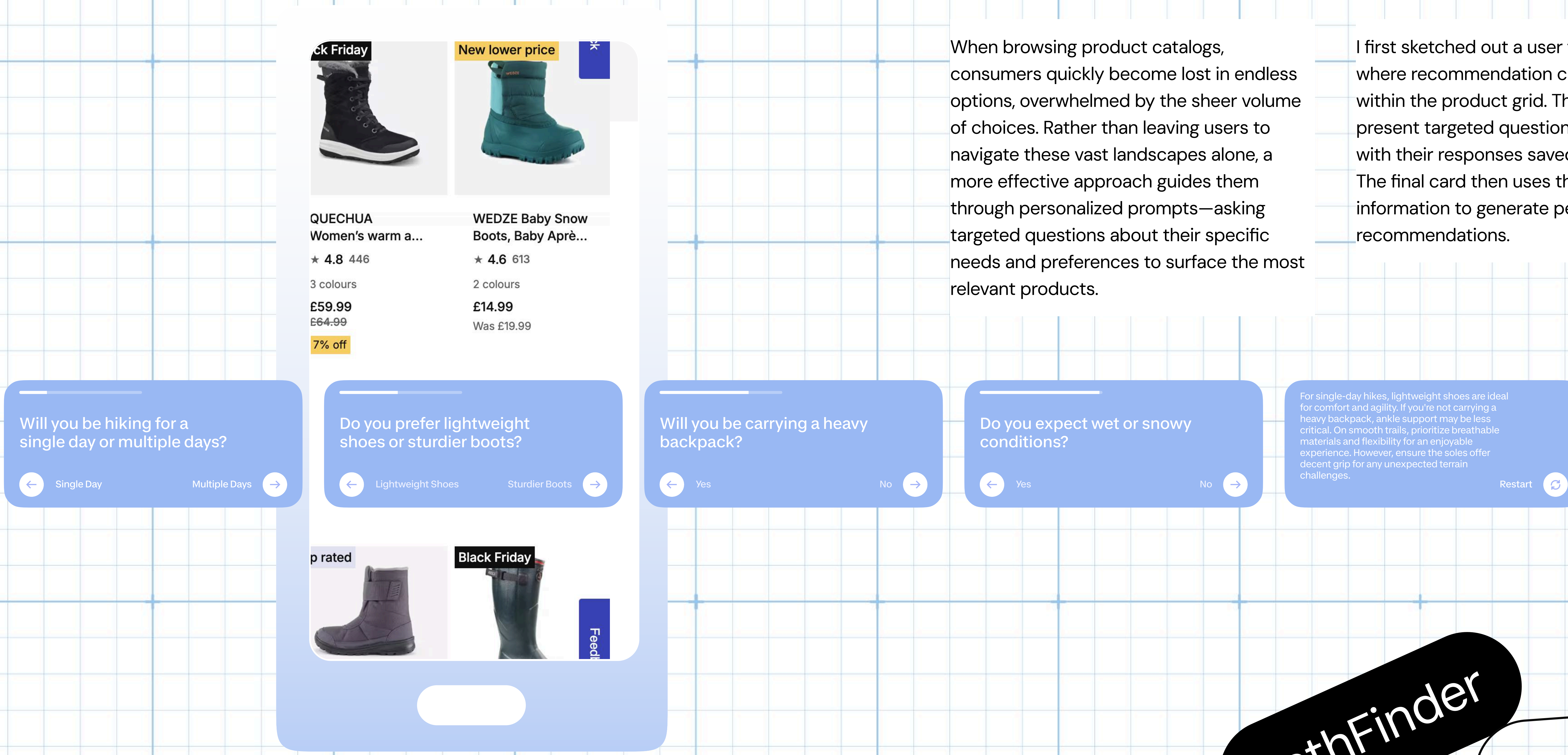
Rapid prototyping isn't just about building quickly—it's about learning quickly. Instead of developing high-fidelity designs and pixel-perfect interfaces, in this stage, I focused on experimenting with web app development and prompting techniques, which revealed diverse approaches to integrating this technology. Following an evaluation of each prototype's desirability, feasibility, and viability, I will demonstrate which solution emerged as the most promising approach for further development.

REFLECTION POINT

If you're an engineer or programming-savvy reader, you may want to skip this section—the solutions won't excite you architecturally. However, you might still find the progression interesting: my prototypes begin with broad, high-level concepts and gradually become more focused and specific as I developed a better understanding of what's achievable through code.

If you're a designer with a trained eye, forgive the unfinished details in these early interfaces. But I'd also love to invite you into the world I experienced—where ideas grow like living creatures, and the goal becomes challenging how information is presented rather than standardizing it.

Prototype 01



Ideation

When browsing product catalogs, consumers quickly become lost in endless options, overwhelmed by the sheer volume of choices. Rather than leaving users to navigate these vast landscapes alone, a more effective approach guides them through personalized prompts—asking targeted questions about their specific needs and preferences to surface the most relevant products.

Process

I first sketched out a user flow in Figma where recommendation cards appear within the product grid. The first four cards present targeted questions to the user, with their responses saved to user input. The final card then uses this collected information to generate personalized recommendations.

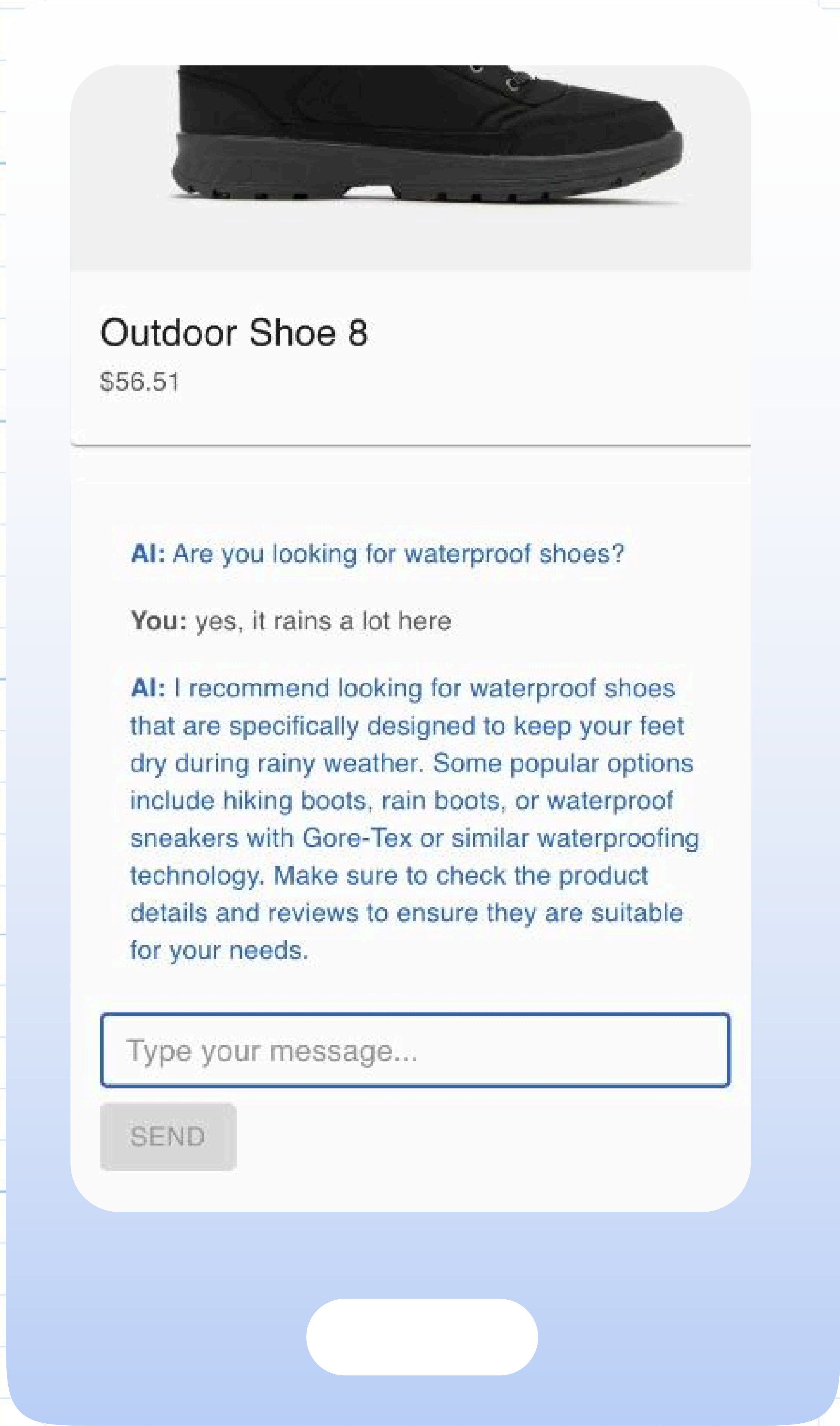
PathFinder

Programming

Then I implemented a React application that integrates the OpenAI API to facilitate the generation of personalized recommendations, using synthetic data as placeholders for the product pages. Below is a code snippet for the prompt sending to the LLM, where the user data is injected into the prompt.

```
const chatCompletion = await
openai.chat.completions.create({
        model: 'gpt-3.5-turbo',
        messages: [
          {
            role: 'system',
            content: 'You are a helpful shoe
recommendation assistant for Decathlon
customers. Generate personalized insights
based on the user\'s responses.'
          },
          {
            role: 'user',
            content:
JSON.stringify(interactions)
          }
        ],
        temperature: 0.7,
      });

      const insights =
chatCompletion.choices[0].message?.content;
```



REFLECTION POINT

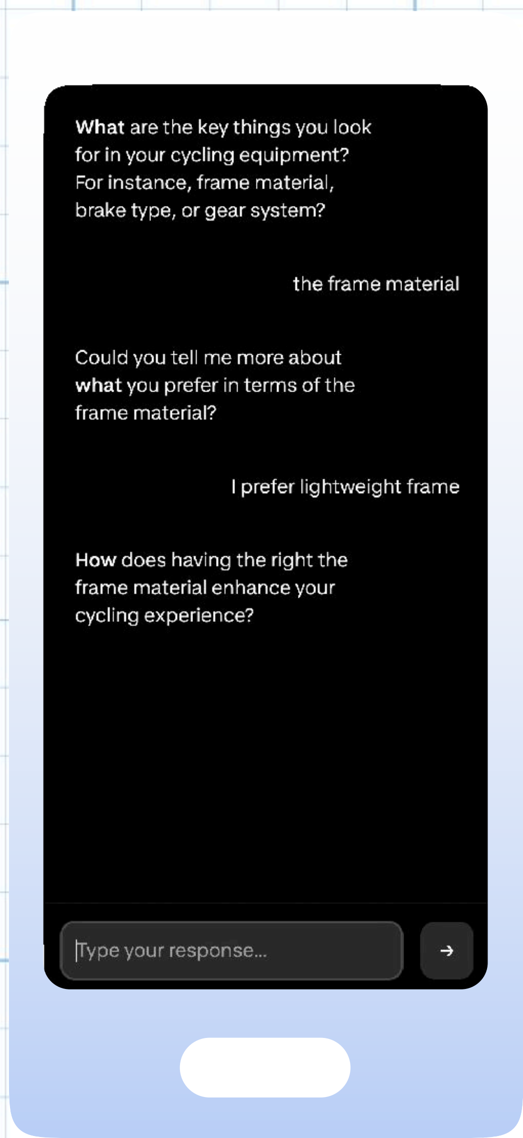
The prototyping took much longer than expected, but looking back, it should have been simpler. What excited me was the possibility of injecting users' contextual information into the experience. However, I discovered a significant challenge: the model's output proved difficult to control, even with clearly defined user context, role, and task parameters. The high temperature setting likely contributed to excessive variation in results, creating an unpredictable user experience that felt entirely new to me as a designer.

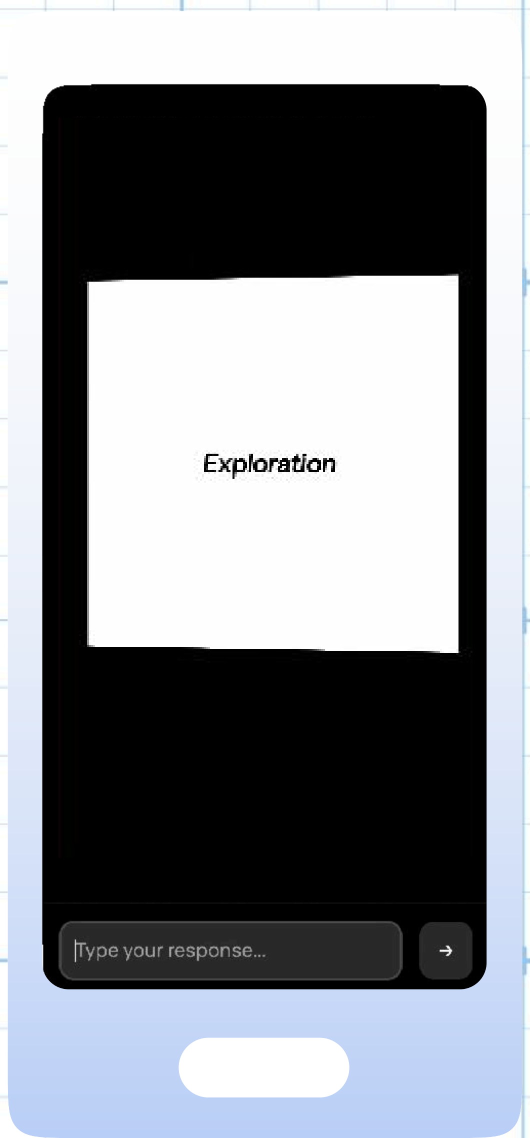
Prototype 02

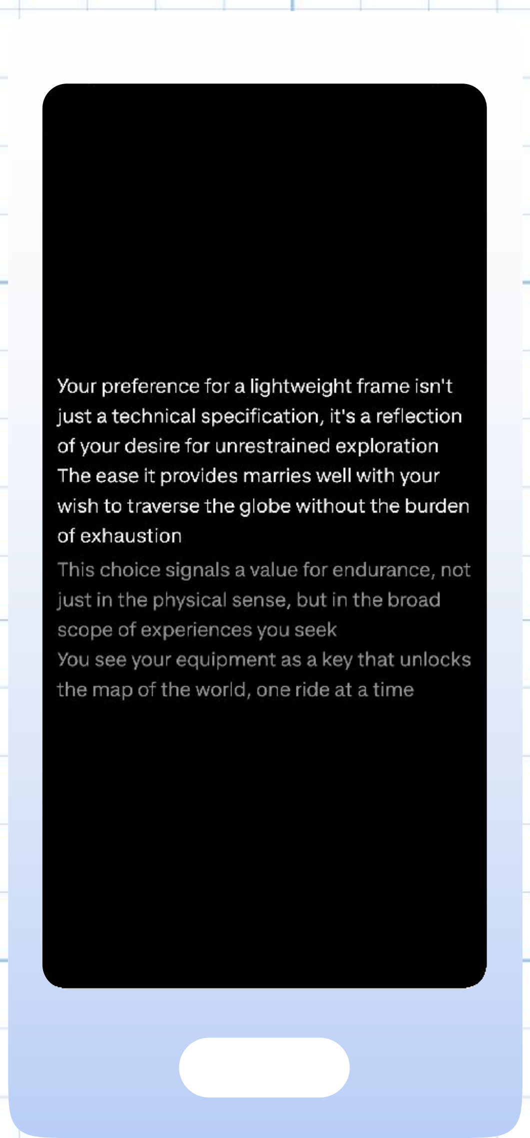
Step1: The user firstly engage in a constructed dialogue

Step2: The system generates a keyword that represents the user's value

Step3: The system generates detailed interpretation and inspire the user







Ideation

In the second experiment, I drew inspiration from the laddering method in interviews. I imagined that through more inviting questions, users could explore their life habits and values more deeply. Here, we could leverage machine intelligence to guide reflection and foster a sense of belonging in sports.

I implemented an interactive dialogue system where users engage in structured conversations. Through carefully designed question sequences, the system encourages users to articulate progressively deeper levels of thought, experience, and emotional connection.

ValueLens

Programming

To implement this functionality, the system first captures and records the conversational exchange as input for the LLM generation process.

Similar to the first prototype, the prompt incorporates user input from this initial dialogue. On the right, you can see the rest of the prompt, where I provided calibrated examples to guide the LLM in interpreting dialogue patterns. This enables the system to extract and articulate the core values underlying user expressions, generating descriptive content that reflects these deeper insights.

```
        messages: [
          {
            role: "system",
            content: You are an empathetic sports coach who excels at
guiding athletes to reflect on their equipment choices and discover
deeper motivations. Your role is to analyze their latest response and
create a reflection that invites growth.

            When analyzing the response, examine:
            1. Word choice and emotion in their language
            2. The connection from feature through benefit to personal
reason
            3. The broader context it reveals about their motivations
            4. The underlying values expressed

            Remember to:
            - Focus only on their most recent response
            - Connect their equipment preference to deeper meaning
            - Identify core values revealed by their choices
            - Create a description that flows naturally sentence by
sentence

            Return as JSON:
            {
              "values": ["1-2 core values identified from this
response"],
              "description": "A reflective narrative about their
perspective, written in clear, separate sentences that can be revealed
two at a time. Each sentence should end with a period. The narrative
should flow naturally and build upon previous sentences. Important: the
content should be less than 100 words."
            }
          }
```

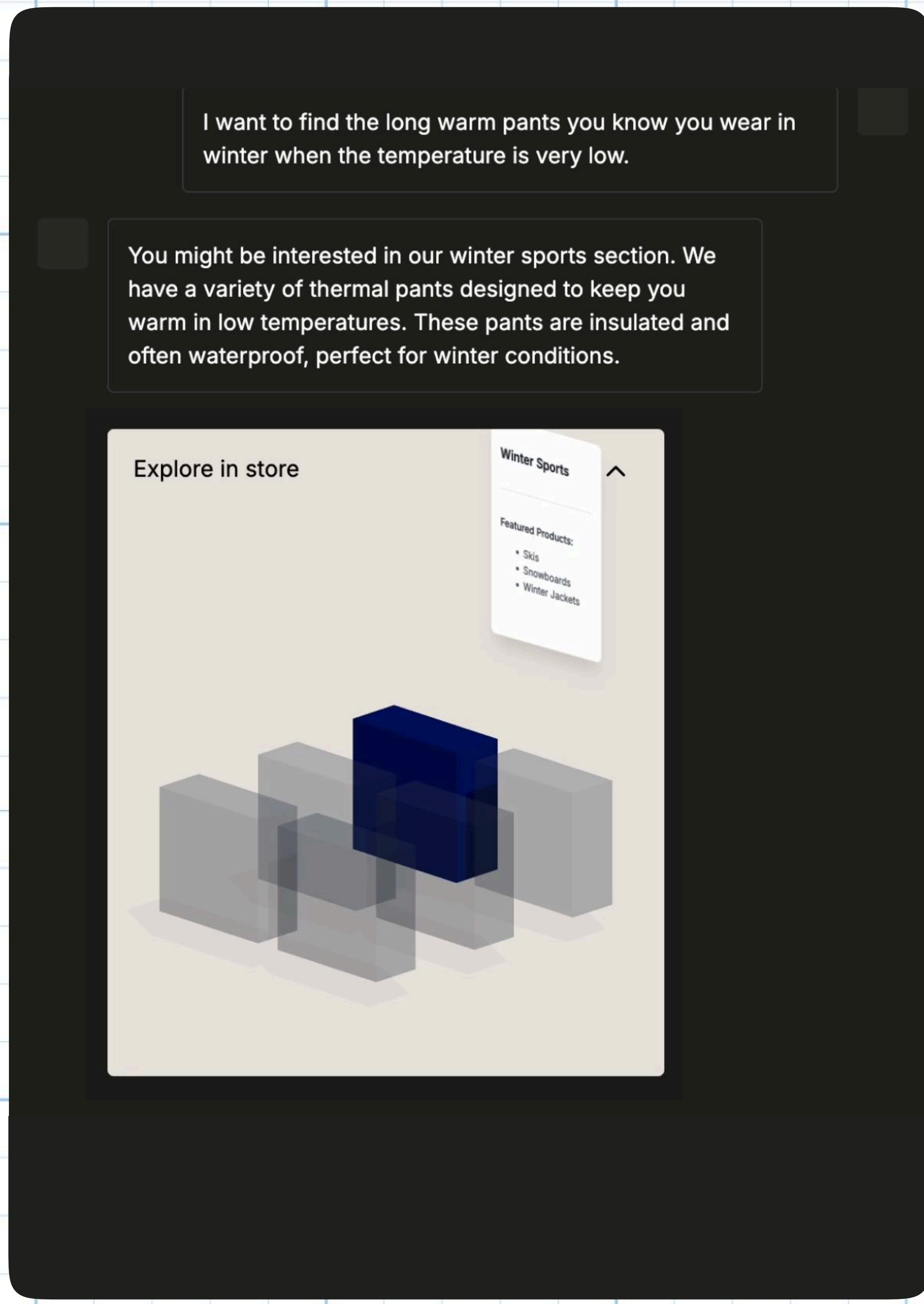
[Prototype 02]
You can access the prototype by scanning this QR code. Permission is required.

REFLECTION POINT

One standout learning from developing this prototype was the genuine fun I experienced with the card flip motion that 'reveals' the generated keyword after the dialogue. The waiting felt meaningful—like anticipating something that would truly connect to what I had expressed. I was curious about how others would react, but this remained unclear since I didn't test with enough people.

The potential I see in this prototype is less functional and more about creating emotional experiences—which has actually deviated from the initial problem of helping users navigate product choices.

Prototype 03



Ideation

In this experiment, I implemented pre-visit spatial orientation through 3D visualization using Three.js. I envisioned that this feature allows shoppers to preview the store layout before visiting, understand product organization, and plan navigation paths. The visualization serves as a cognitive mapping tool, helping users develop spatial awareness before entering the physical space. This approach addresses navigation challenges by providing spatial knowledge that reduces disorientation and inefficient movement within the retail environment.

SpaceMap

Programming

To implement this feature, I integrated an LLM that analyses user interests and identifies relevant store sections, as shown in the code demonstration. The system then highlights these personalised areas within the interactive 3D map, creating a customised spatial guidance experience.

```
    const prompt = `
      You are a sports expert. We have
these categories:
      1) winter sports
      2) hiking
      3) cycling
      4) water sports
      5) team sports
      6) fitness

      The user asked: "${userText}"

      INSTRUCTIONS:
        Return only JSON. No extra text, no
code fences, no explanation outside JSON.
        The JSON must have the format:
        {
          "sections": [1, 2],
          "answer": "some short guidance
for the user"
        }

        Example 1
        User says: “I love skiing and
surfing.”
        Desired JSON: {"sections":[1,4],
"answer":"Short guidance ..."}

      ....
```

REFLECTION POINT

Compared to the previous prototype, I took this a step further by experimenting with LLM output integration. Instead of displaying the response on screen as a traditional "answer" to a "question," I used the LLM output to drive the 3D interface itself.

While it's unclear whether this workflow outperforms rule-based detection of user input, it revealed the potential of using LLMs in novel ways—restructuring output data to create more user-friendly experiences.

[Prototype 03]
You can access the prototype by scanning this QR code. Permission is required.

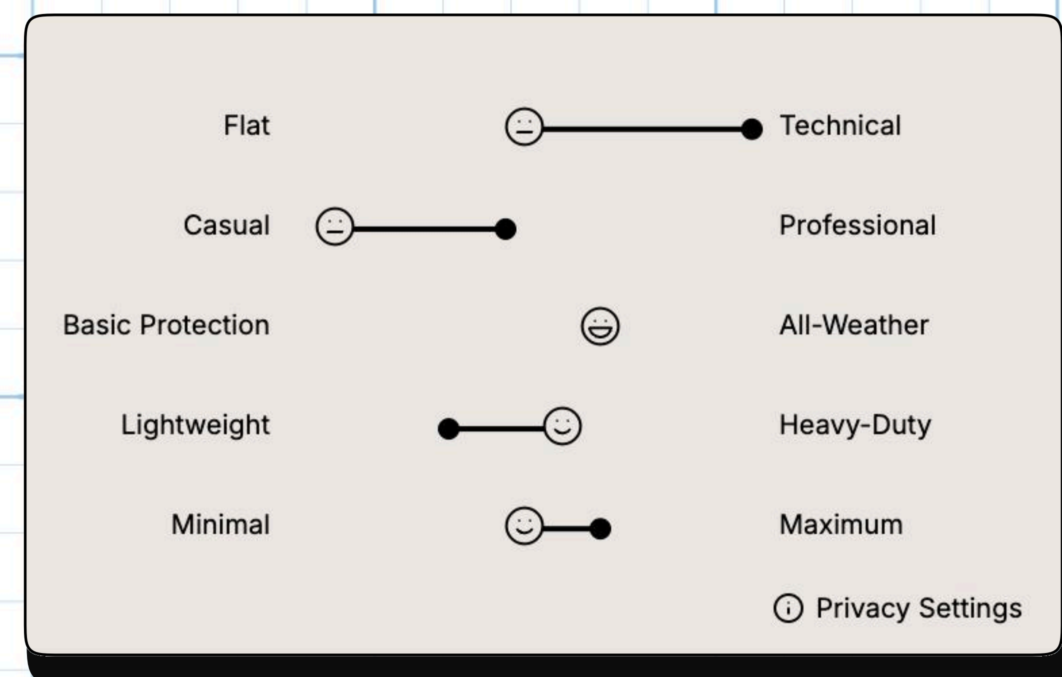
Prototype 04

Ideation

In this experiment, I explored using LLMs to streamline product evaluation process by analysing items along customised dimensions tailored to specific products and individual user preferences.

The system presents compatibility assessments through intuitive visual cues like distance indicators or expressive emojis, allowing users to quickly identify where products align with or diverge from their ideal criteria.

This prototype was only partially implemented. The mock-up on the right was used as a communication tool. A revised version of this design will be discussed later on.



DimensionMatch

Decision Matrix

Expert evaluation of LLM response variations

As the experimentation phase progressed toward convergence, I employed the widely recognised evaluation framework of 'desirability, feasibility, and viability'—a standard approach for assessing design concepts, particularly in organisational contexts (Gonen, Esra 2019). This framework served to evaluate and prioritise which concepts warranted further development resources.

Prototypes	Feasibility	Desirability	viability
PATHFINDER	–	+	++
VALUELENS	+	–	+
SPACEMAP	–	++	+
DIMENSIONMATCH	++	+	+

As illustrated in table, several limitations emerged across the prototypes. The *PathFinder* system, while promising, presented feasibility challenges due to the complexity of adapting it to diverse product pages within my project scope. The ValueLens feature scored relatively low on desirability, failing to demonstrate clear business potential. Meanwhile, the SpaceMap, despite showing promise in helping consumers plan visits, proved infeasible to implement with a proper database architecture within the constraints of this project.

In conclusion, following the exploration phase, I selected DimensionMatch as the final prototype for further development based on its alignment with the challenges identified in the problem definition. This choice was particularly motivated by the prototype's strengths in two critical areas: interpreting diverse language patterns through personalized dimensional analysis of products, and mitigating information overload by extracting and presenting relevant data in a streamlined format.

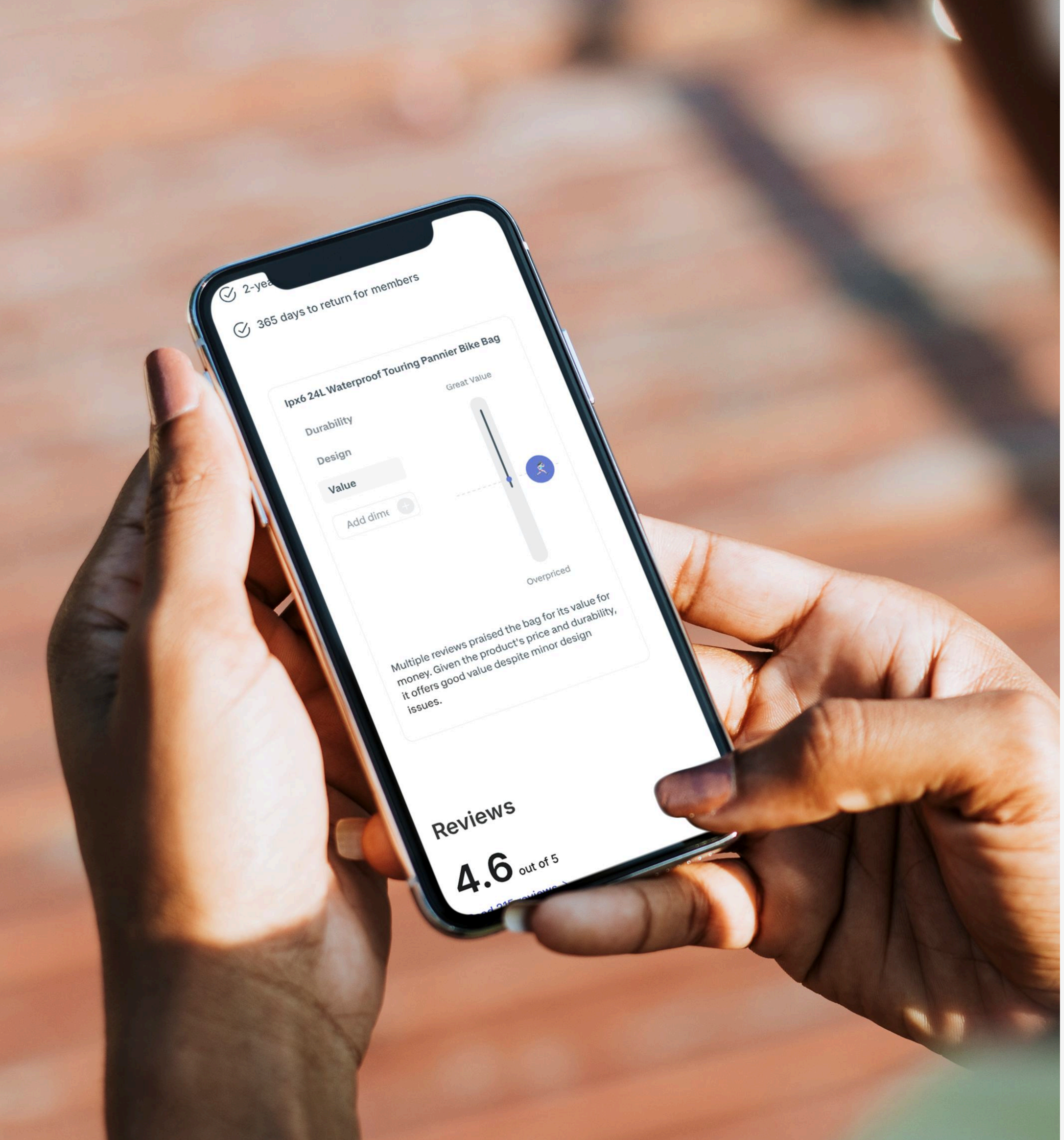
REFLECTION POINT

The evaluation feels somewhat rushed and, once again, lacks collective reflection on the prototypes. Ideally, expert validation would have been invaluable at this stage—identifying the strengths and limitations of each prototype and assessing how effectively each approach might help consumers navigate product choices.

Further developing the prototype for better integration

Through the exploration of a variety of ideas, several promising functions were developed, with DimensionMatch emerging as particularly significant. However, these innovations have not yet been integrated into the current Decathlon website architecture. Therefore, in the next development phase, the design goal was established to reimagine the existing website design with a comprehensive integration of DimensionMatch, ensuring its functionality enhances rather than disrupts the user experience.

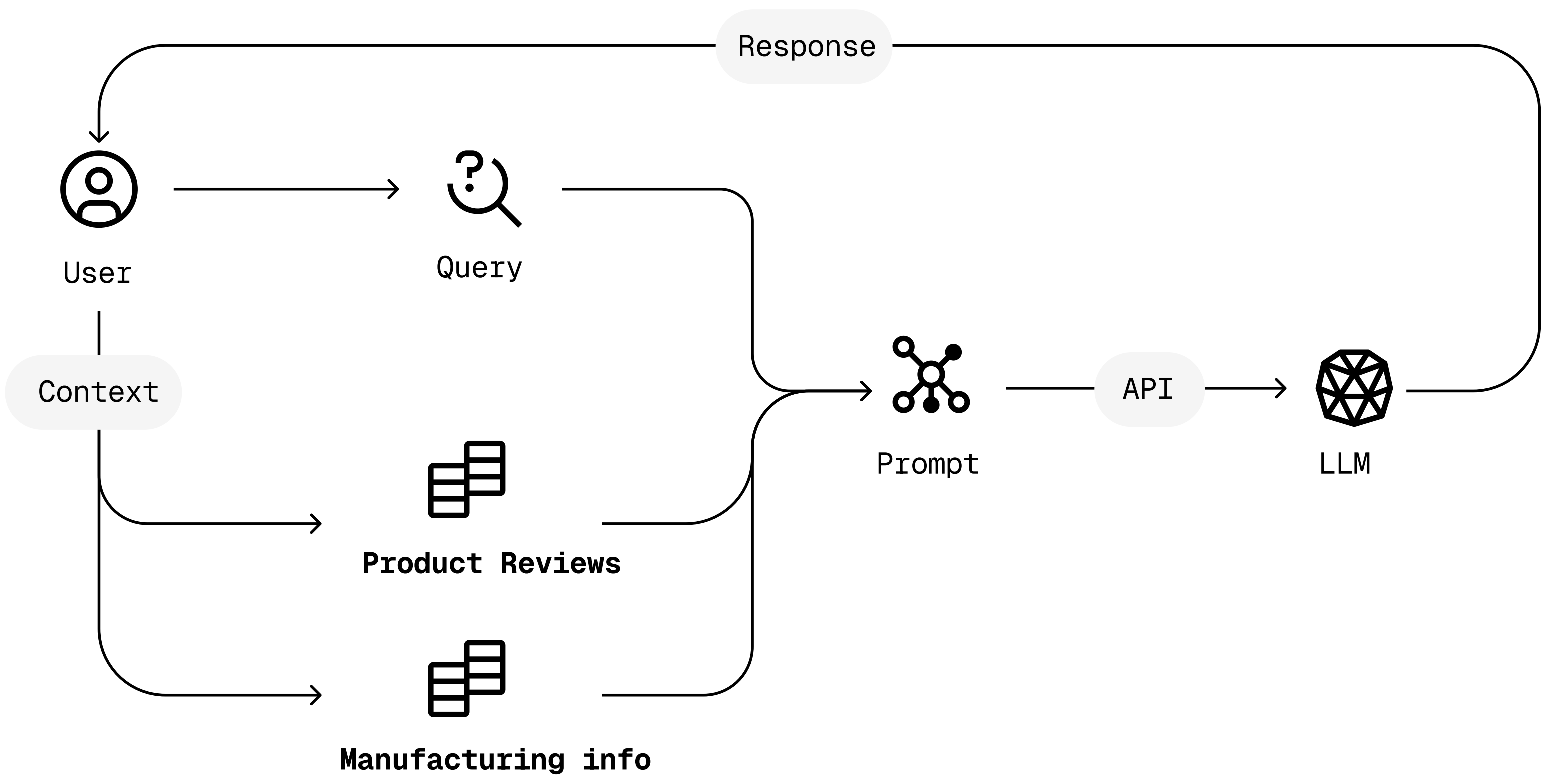
During this phase, several key design decisions were made to address the project scope, environment constraints, and user testing requirements. Compared to the creative diverging phase, this stage moved toward convergence to prepare for a development cycle and gain deeper understanding of users' opinions.



The concept was implemented into an interface that follows the design language of the current website and integrated into the product page. This choice became clear when considering its function which is analyzing the match between products and users.

As presented on the right, the dimensions are listed on the left of the screen, allowing users to switch between them and view the corresponding values on the right on the right of the screen. This design change saves space compared to the previous prototype, for better integration into the existing interface.

An additional design consideration was to provide justification for the match score underneath. One design choice made was to use user reviews data for synthesizing and generating the dimensional score and justification.



To incorporate the concept, on the technical level, we need to embed the relevant knowledge into the LLM system. To do that, we could use modern techniques like Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), in which pipelines the user queries will be processed before getting into the LLM. With a searching phase that uses vector or semantic search, looking for relevant information, the relevant information will be put together with the original query into the model.

Due to time constraints, a technically simpler yet less cost-efficient technique was used, where a simpler traditional search algorithm was implemented. In this approach, the user's query, the context (in this case the product review and manufacturing information of the product on the current page), along with other prompting techniques were combined into a single prompt for each request.

[Prototype 05]
You can access the prototype by scanning this QR code. Permission is required.

User Testing: Early Validations

A good practice when developing design is to conduct early and iterative user testing: test, fix, and test again (Krug & Black, 2009). With that spirit, I conducted small-scale usability testing before pursuing systematic user studies.

During the first usability test, the participant was first introduced to the core functionality and then explored the redesigned interface with the designed component implemented on a backpack product page. The decision to use a backpack was made due to its gender-neutral nature, making it easier to recruit diverse participants for testing. Throughout the process, the participant was encouraged to think aloud, and the session was recorded with a transcript generated for analysis purposes.

There were several validation points and points of improvement on the usability level. In particular, the user expressed a strong positive reaction to the summarization functionality, commenting that -

> **“22 pages (of reviews) is just too much to read”**

When seeing relevant information extracted, such as the fact that **'it could fit a 13-inch laptop,'** the user was notably impressed. The user further suggested that

> **“I hope AI could also ask me what things I want to store in my bag and predict which bag suits me better, otherwise I need to analyze this information in the user reviews myself.”**

On the usability level, the visualization seemed to be not intuitive, and the user questioned the criteria chosen for visual analysis:

> **“I don't know why awkward is the opposite of intuitive?”**

Additionally, it was questionable if it is the best design choice to use a limited input area and show examples of keyword dimensions. The user commented that

> **“ It would be nice to have better guidance here from usability perspective, for example, if in the first three examples, you have one that is longer it will be more intuitive.”**

WWWWWH Analysis: Guiding the Next Design Iteration

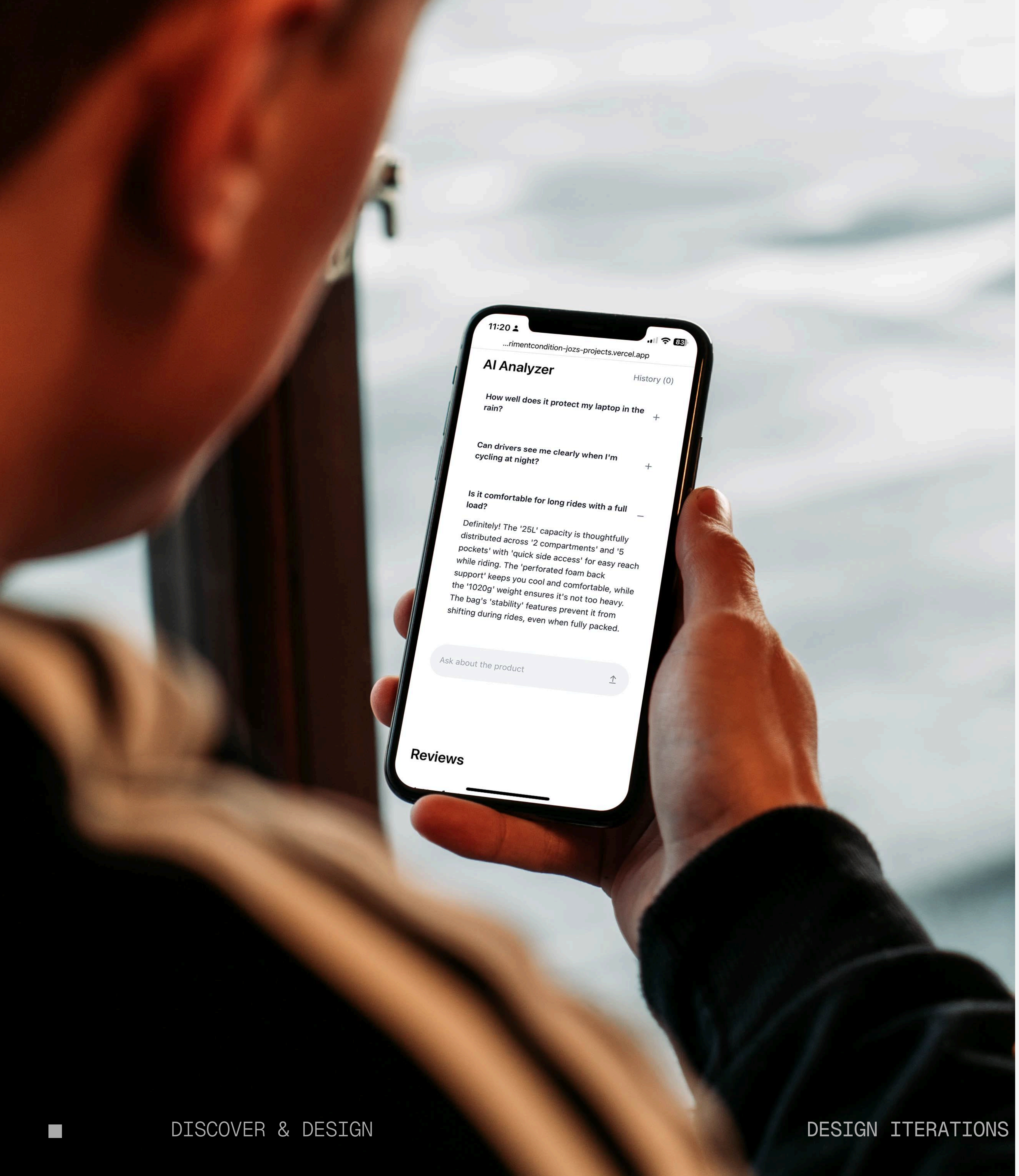
Reflecting on our usability tests and project environment, I realized I need to sharpen my focus when testing and designing features. Building upon the WWWWWH framework (Van Boeijen et al., 2020), I'm expanding my design considerations to include:

- What: What other essential information should be presented in the interaction?
- Where: Where should components be located for optimal visibility and flow? Where do users expect to find these features in relation to their mental models?
- When: When does the component become relevant in the user journey? When should information appear or disappear to create a coherent experience?
- How: How should information be visually presented? How can interactions be made more intuitive?
- Why: Why should users care about this information?
- Who: Who are the primary users of these features?

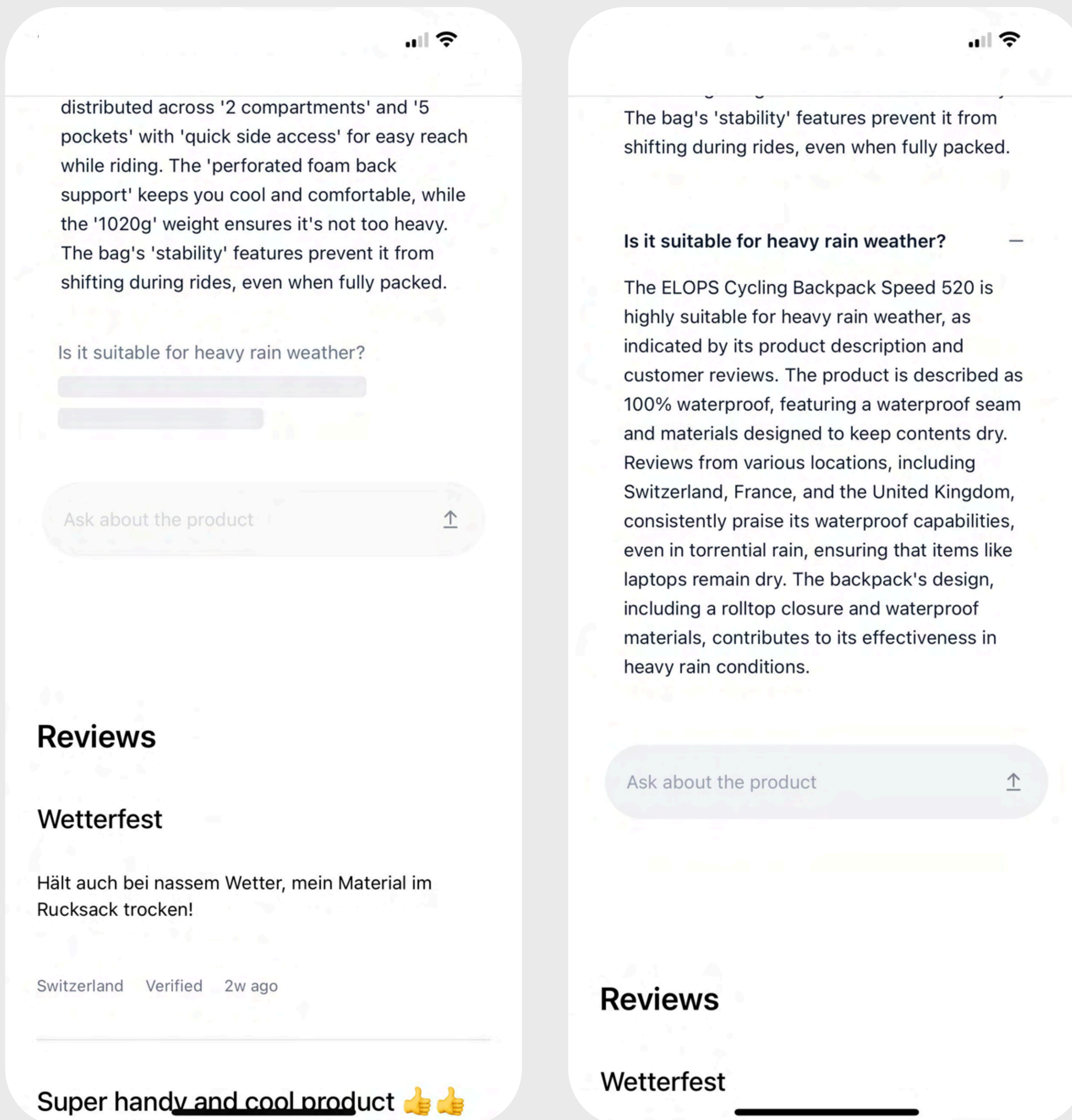
As discussed in the related literature, LLM-human interactions often struggle in two areas: they fail to adapt to users' intuitive spatial memory and provide limited support for users developing coherent mental models of the system. These challenges are directly linked to the fundamental questions of **when** and **where** information should be presented to serve the higher goals of comprehension and informed judgment. To better understand these dimensions, I zoomed out in the next design iteration and contextualised the design within the actual shopping user journey.

Additionally, the **What** question presents significant value in exploring the core utility of LLM-generated responses, as it addresses the inherent purpose and promise of the technology itself.

Finally, the innovative approach of integrating **user reviews** through LLM technology merits deeper investigation through additional user testing to thoroughly understand user reactions, perceptions, and the overall effectiveness of this design strategy.

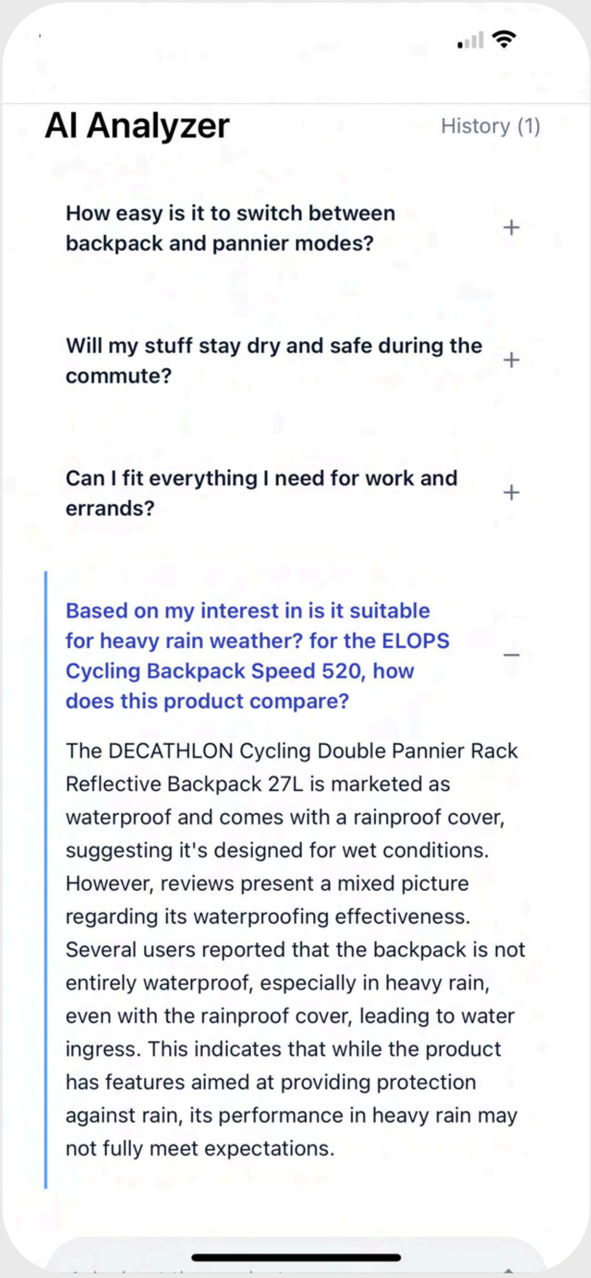


For this iteration, I chose to adapt a traditional FAQ design pattern to display example questions, replacing simple keywords with more complete, natural language expressions. Below these examples, users find an input bar where they can ask their own questions. When a user submits a query, the system generates a response and then adds this new question-answer pair to the growing list of interactions (as shown in the next page), creating a dynamic, personalized FAQ that evolves throughout the user's session.



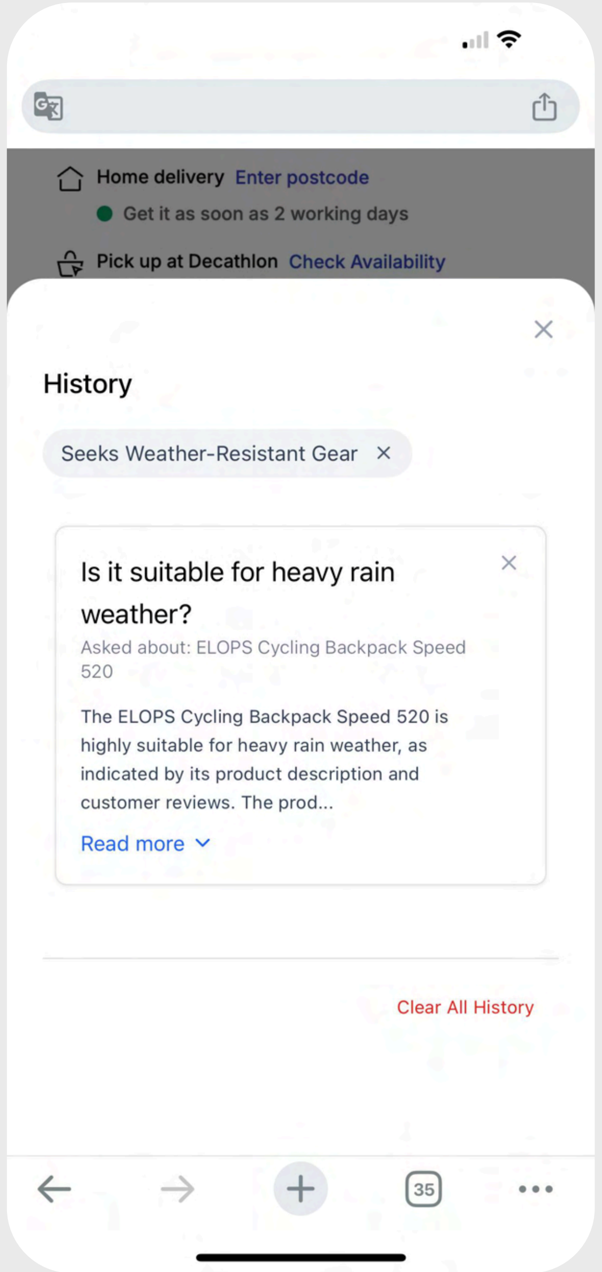
Generated Response

As demonstrated in the screenshot of live demo application, the response synthesizes multiple information sources into a piece of text, blending product specifications (such as material descriptions and design details) with user reviews (such as geographic distribution of reviewers and collective opinions).



Context sensitivity

A cross-product memory feature was implemented that preserves the context of previous inquiries as users navigate between products. When a user moves to a new product page, the system transfers their previous question and prompts the LLM to generate a comparable response addressing the same concern for the current product.



Transparency

Additionally, the system maintains a comprehensive record of all questions asked in a dedicated overlay panel, providing transparency about what information the system has collected about the user.

Reframing The Problem

Through the process, a concept emerged that consolidated multiple assumptions. The study focus narrowed from exploring design patterns for user navigation to validating a specific concept: embedding an adaptive component within product pages that enables users to explore product information more intuitively, thus gaining better understanding with less effort compared to reading "22 pages of reviews”.

Moving down to the usability level and examining specific design decisions more closely, the interest centered on understanding how the FAQ pattern, context sensitivity and preservation, and the history panel would impact user experience. This experience can be evaluated through both pre-experience expectations and post-experience perceptions—both crucial elements for success. When users first encounter the interface, they should intuitively understand how to use it; after using it, the execution should fulfill their needs. This relates to Norman's concept of the gulf of evaluation (Norman, 1988), which describes how effectively a system presents information that users can readily understand and interpret in relation to their expectations and goals.

This prototype not only addresses these usability considerations but also serves as a bridge between our research questions. While our initial focus was on how LLM technology could improve navigation for Decathlon website users, this implementation allows us to explore more nuanced questions about how non-conversational LLM interaction patterns compare to purely conversational interfaces in terms of both user experience and expectation management. Through this progression, we're building a foundation to understand which interaction models best serve users in product exploration contexts.

Hypothesis

RESEARCH QUESTION 1

How might we leverage LLM technology to improve navigation experience for decathlon website users?

Hypothesise that integrating contextual LLM capabilities directly into product pages will reduce cognitive load and increase information discovery efficiency compared to traditional browsing approach.

RESEARCH QUESTION 2

How do non-conversational LLM interaction patterns affect user experience compared to purely conversational interfaces?

Hypothesize that structured, non-conversational patterns like the FAQ approach will lead to higher user satisfaction for product-specific inquiries.

RESEARCH QUESTION 3

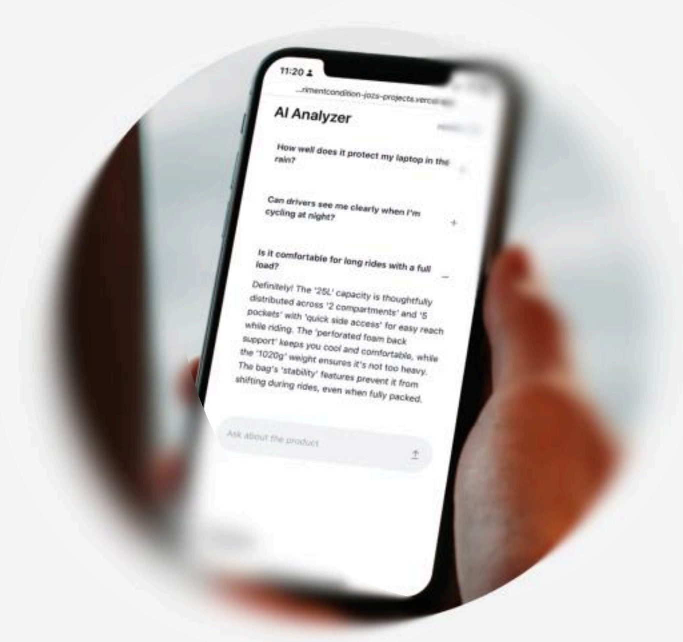
How do non-conversational LLM interaction patterns manage user expectations compared to purely conversational interfaces?

Hypothesize that more constrained interaction patterns will create clearer mental models for users, resulting in fewer misunderstandings about system capabilities and more accurate predictions of system behavior.

Static Webpage

RQ 1

Does it solve the problem?

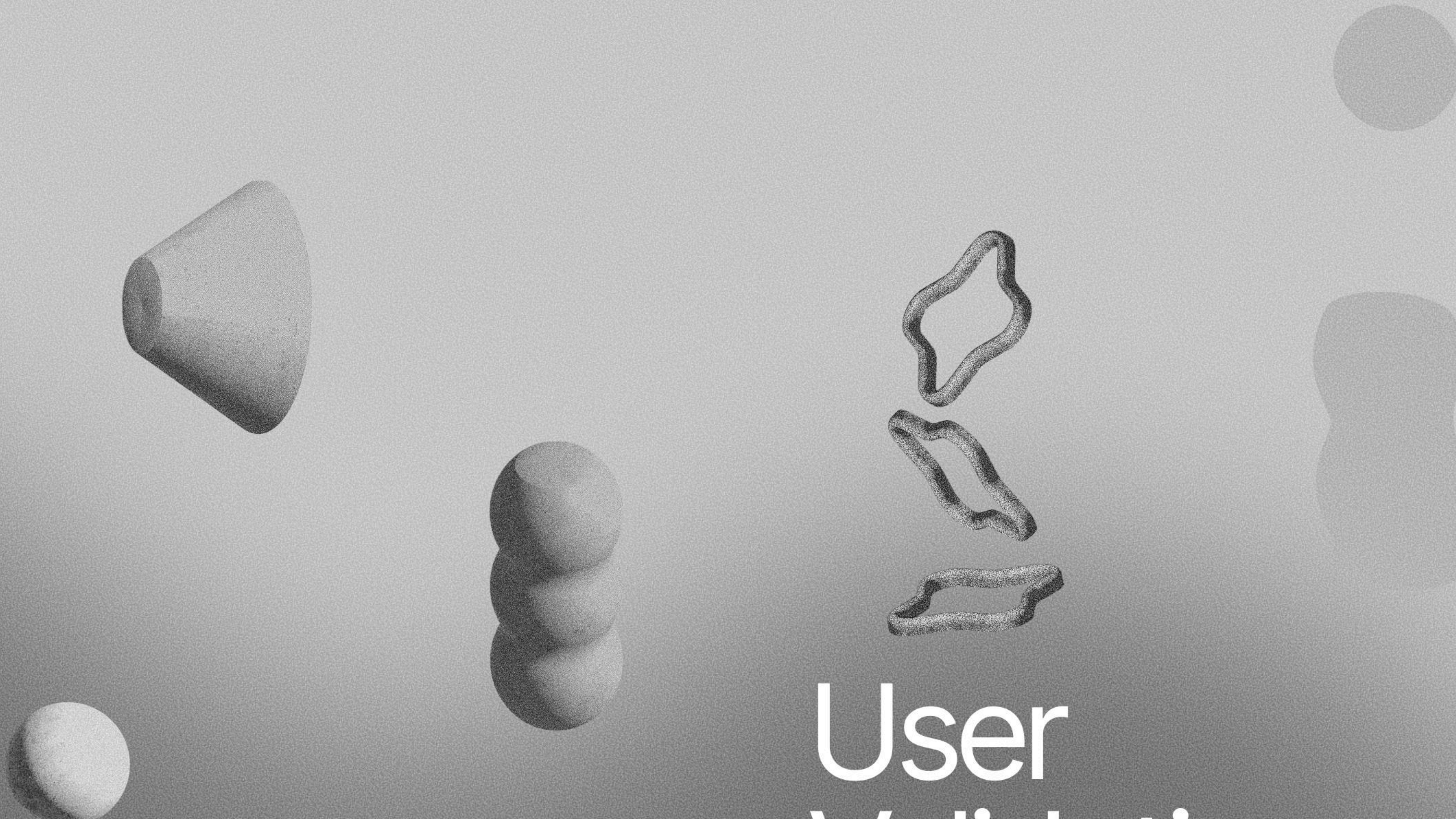


RQ 2

RQ 3

How does it compare?

Chat-based Interface



User Validation

4.1

User Study Method & Preparation

Pilot study, additional interface design and other preparation before user study

4.2

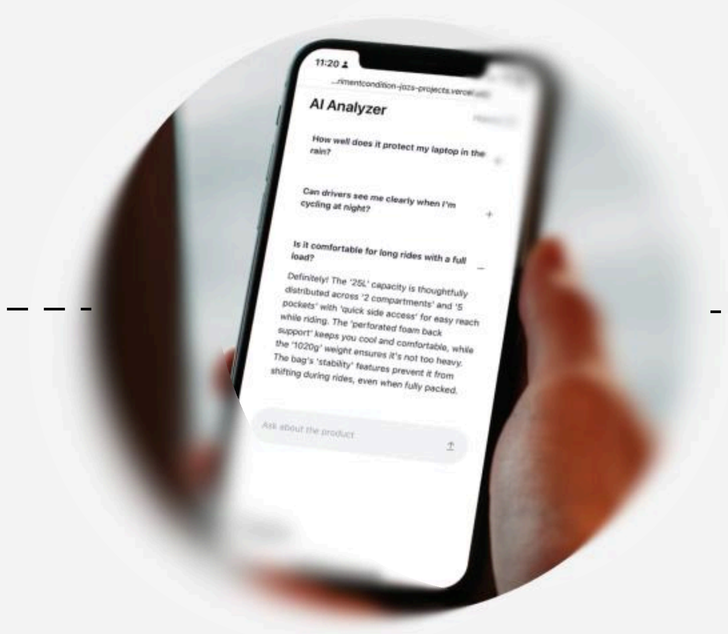
Data Collection and Analysis

Showcase the insights from the user study in both textual and visual format

Learn from users: Validation & Discovery

This chapter analyzes feedback collected from diverse users and identifies meaningful patterns that emerged. This process serves a dual purpose of validation and discovery – evaluating how effectively the prototype addresses the identified problems through quantitative measurement, while also examining how user testing has enriched understanding of designing human-LLM interactions in this specific context. Building on the research questions about leveraging LLM technology for navigation experience, comparing conversational and non-conversational interaction patterns, and managing user expectations, the analysis reveals whether the hypotheses hold true.

Static Webpage

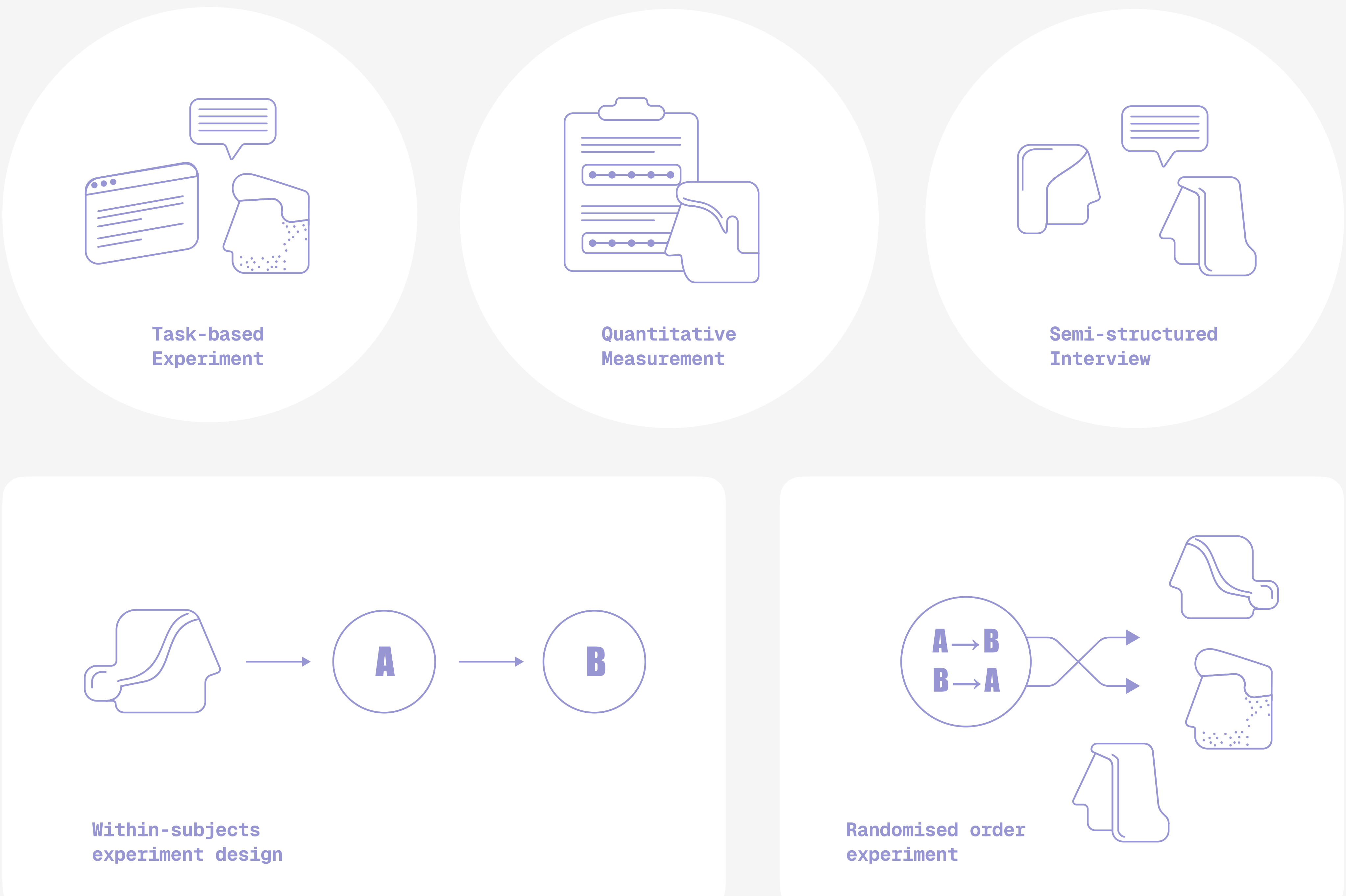


Chat-based Interface

Qualitative Validation

Quantitative Validation
Qualitative Discovery

Methodology



With the rich objectives, the study demands not only comparative analysis but a comprehensive methodological framework. To address both validation requirements and discovery objectives, I designed a sequential mixed-methods approach (Creswell & Creswell, 2017).

This integrated methodology combines quantitative measurements with qualitative investigation through think-aloud protocols and semi-structured interviews, ensuring methodological rigour while capturing rich experiential data. This approach allows me to elaborate and expand the findings from one method with insights from another, providing the flexibility to accommodate each participant's different emphases and perspectives.

Another crucial consideration is our objective to validate the design approach by comparing it with the traditional or existing interface—a task requiring careful experimental design.

Given the objective to understand the differences between the two conditions, combined with time constraints limiting extensive participant recruiting and experiments conduction, a within-subjects design was chosen. To mitigate potential order effects, the two conditions were randomised across participants, ensuring methodological integrity while maximising the insights gained from each participant interaction.

Pilot Study & Adjustments

To assess the effectiveness of the experimental design and refine interface details, a pilot study was conducted with three students from the Industrial Design Engineering Faculty. The study compared the control condition (the existing product page on the Decathlon website) with the experimental condition (our newly designed interface).

For the task design, we tested a use case that effectively guided users to explore the interface. The scenario involved helping a friend select from among three product options based on specific contextual requirements: "Reliable, practical, easy to use; Suitable for weather in the Netherlands; Environmentally conscious." This design choice aimed to simulate conditions closely resembling a real shopping experience, enhancing ecological validity while maintaining experimental control.

For the quantitative assessment, our initial questions focused on single measures: confidence in judgment, clarity of information presentation, and overall satisfaction. However, this approach yielded limited data for comprehensive analysis, prompting further investigation into established user testing dimensions. The work from Kocielnik et al. (2019) was identified as a significant reference with measurement frameworks, based on its similar objectives and objectives with this study. One aspect of the measurement was focusing on user expectations, which builds on the Expectations Confirmation Model. This framework includes pre-intervention measures, in this study was adapted to "How well do you expect the AI Assistant to work?", and post-intervention assessment ”How well do you feel the AI Assistant works?” Both were answerable on a scale from 0% to 100% with 10% increments.

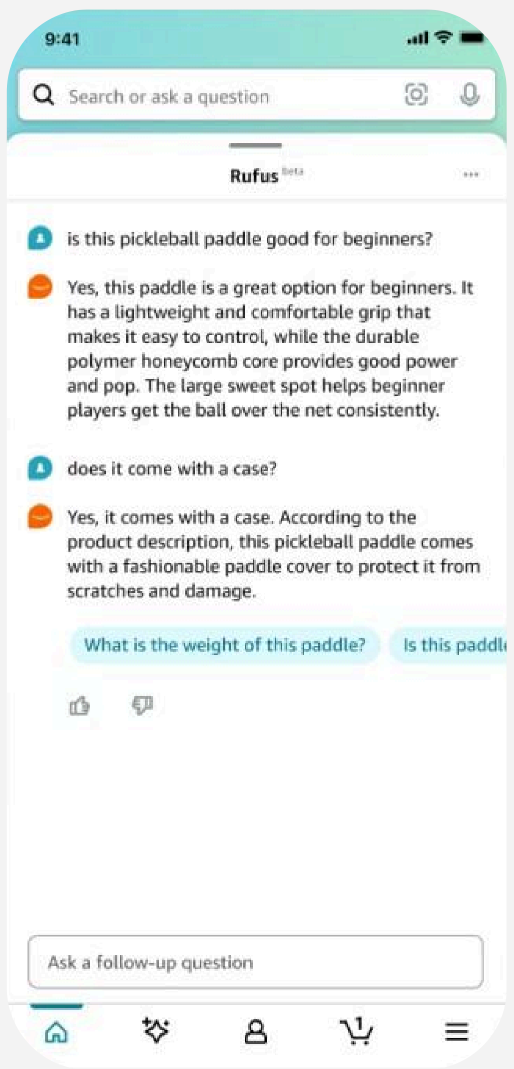
Measurements were collected at two specific points: first after users were exposed to the static interface showing a working status, and subsequently following their engagement with the full features through the designed task. This two-phase measurement approach allowed us to capture both initial impressions and informed assessments after meaningful interaction.

Both pre-intervention and post-intervention assessments were conducted across four identified dimensions. The adaptation focused specifically on these dimensions as they represent criteria particularly relevant to our use scenario.

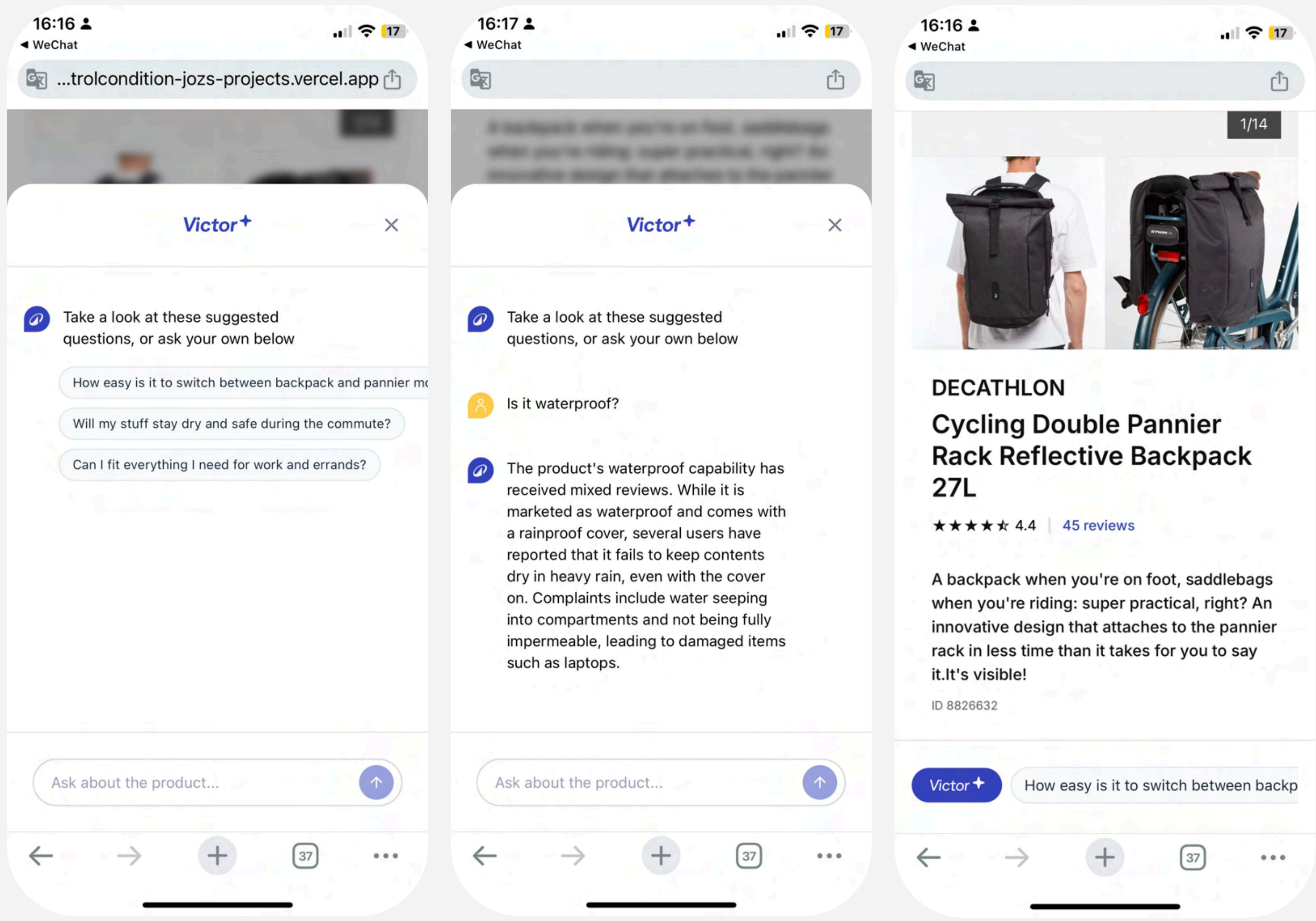
The second quantitative measurement framework addressed overall satisfaction, which was adapted from the same paper, based on the Technology Acceptance Model (TAM). This measured satisfaction across five dimensions: future use, recommendation to others, helpfulness, productivity, and annoyance, answerable with a 5-point Likert scale.
Lastly, the understanding of the AI component was measured through the question "I feel like I have a good understanding of how the AI will process my requests," answerable with a 7-point Likert scale.

For the two design conditions, adaptations were implemented following insights gained from the pilot study. To establish a meaningful comparison between our designed interface and a chat interface, we selected Amazon's Rufus as our benchmark for the control condition, given its comparable functionality in delivering product information.

As illustrated in the right panel, the control condition incorporates several hallmarks of state-of-the-art chat interfaces: a personified name; a structured chat interface featuring distinctive icons representing system and user, alternating throughout the interaction sequence; and a selection of suggested questions to facilitate user engagement and guide interaction.



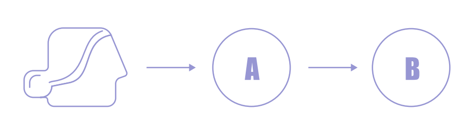
Benchmark
Amazon Rufus



Design
Control Condition

Final Study Protocol

Break-down of the experiment



(Demonstrated with the example of first receiving prototype A, then receiving prototype B)

The study design was extended from the pilot study design without major changes except for the design conditions change. The study was conducted either virtually through online conference platforms like Teams or Google Meet, or in-person at the Industrial Design Engineering faculty. In both settings, participants were informed about consent procedures, which included screen recording of prototype interactions and audio recording with automatic transcription of interviews.

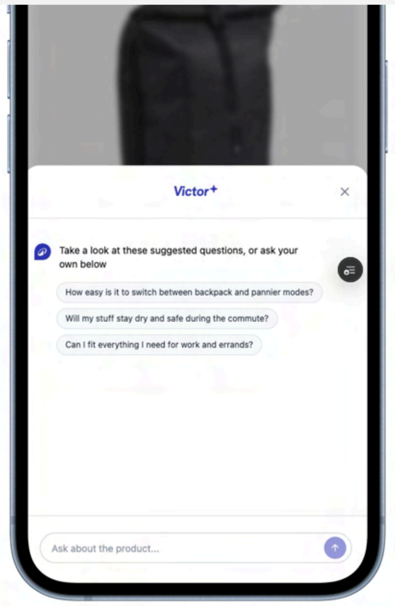
For the A/B testing phase, each participant first received a Qualtrics survey containing pre-task measurement questions, followed by guidance for the use case and a link to the prototype. While completing the assigned tasks, participants were encouraged to think aloud, verbalizing their thoughts and actions. After completing the tasks, they filled in post-task measurement questions. This entire process was then repeated with the second prototype assigned to each participant.

Finally, participants shared their experiences with both prototypes in an interview that specifically explored the differences they noticed between the two design approaches.

1 Pre-task Measurement

2 The 1st Task

3 Post-task Measurement



How well do you expect the AI Assistant to work?
(11 points scale)

I feel like I have a good understanding of how the AI will process my requests.
(7 points scale)

Your friend Alex just moved to the Netherlands and asked you for help finding a backpack suitable for these specific requirements:

- Reliable, practical, easy to use
- Suitable for weather in the Netherlands
- Environmentally conscious

Please help Alex compare these two products based on the requirements

[Experiment]
You can access the prototype by scanning this QR code. Permission is required.

How well do you feel the Victor works? (11 points scale)

I am satisfied with how well the Victor worked (5 points scale)

I would use the Victor if it was available. (5 points scale)

I would recommend the Victor to my friends and colleagues. (5 points scale)

I found the Victor to be helpful. (5 points scale)

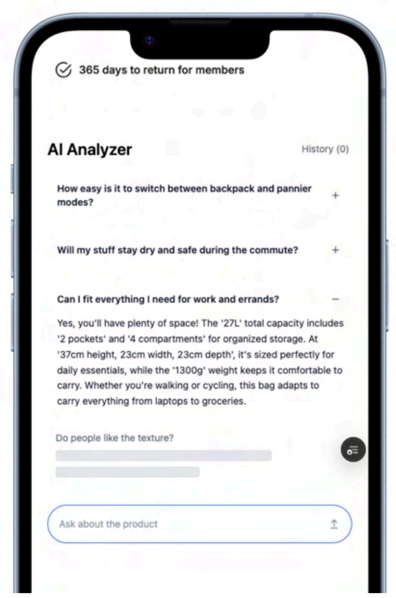
I found the Victor to be annoying or distracting. (5 points scale)

4 Pre-task Measurement

5 The 2nd Task

6 Post-task Measurement

7 Semi-structured Interview



How well do you expect the AI Assistant to work?
(11 points scale)

I feel like I have a good understanding of how the AI will process my requests.
(7 points scale)

Your friend Alex just moved to the Netherlands and asked you for help finding a backpack suitable for these specific requirements:

- Reliable, practical, easy to use
- Suitable for weather in the Netherlands
- Environmentally conscious

Please help Alex compare these two products based on the requirements

[Control]
You can access the prototype by scanning this QR code. Permission is required.

How well do you feel the AI Analyzer works? (11 points scale)

I am satisfied with how well the AI Analyzer worked (5 points scale)

I would use the AI Analyzer if it was available. (5 points scale)

I would recommend the AI Analyzer to my friends and colleagues. (5 points scale)

I found the AI Analyzer to be helpful. (5 points scale)

I found the AI Analyzer to be annoying or distracting. (5 points scale)

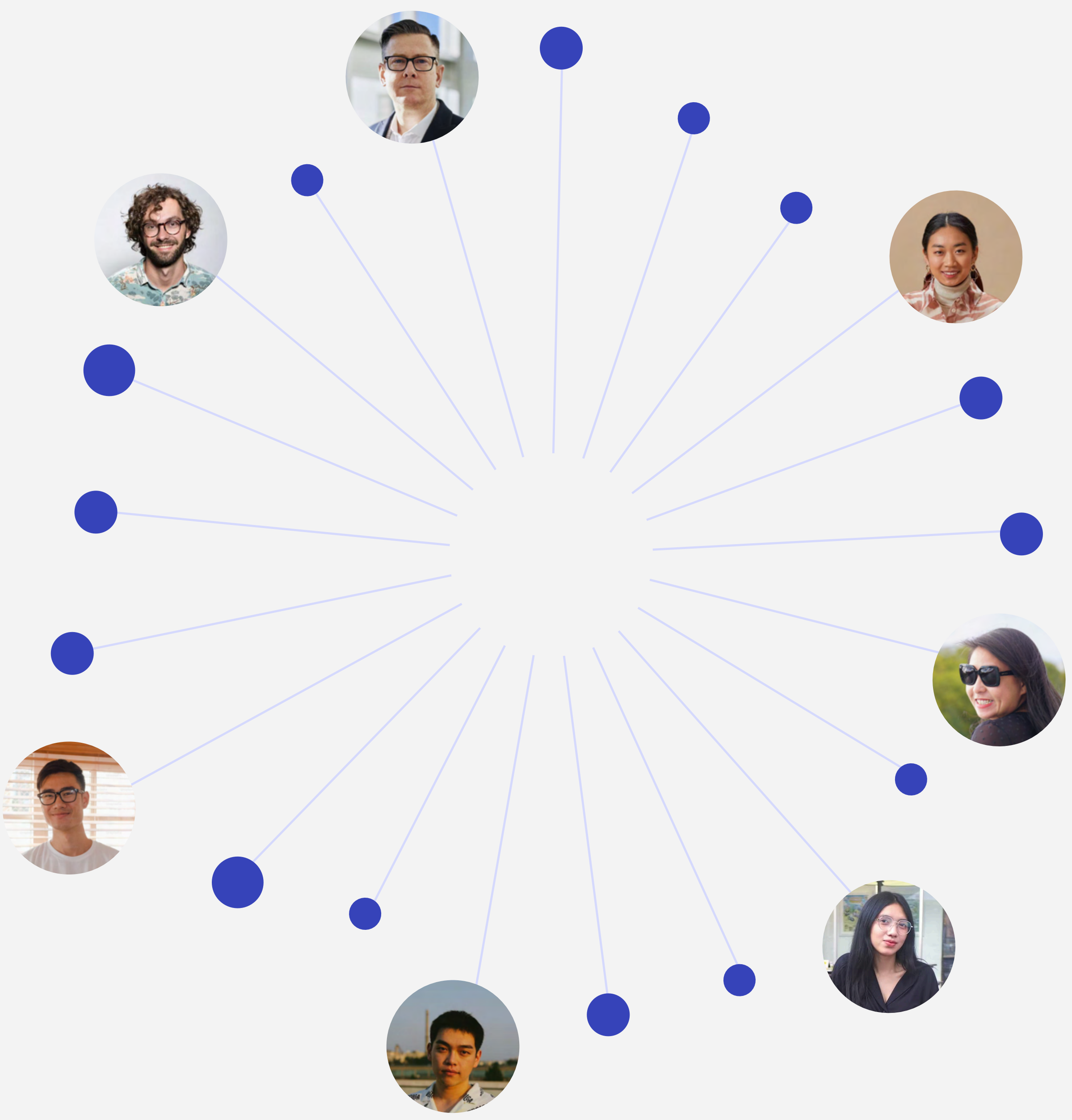
“How would you describe the difference between this two AI products in your own sentences?”

“Were there moments when either AI surprised you - either positively or negatively? ....

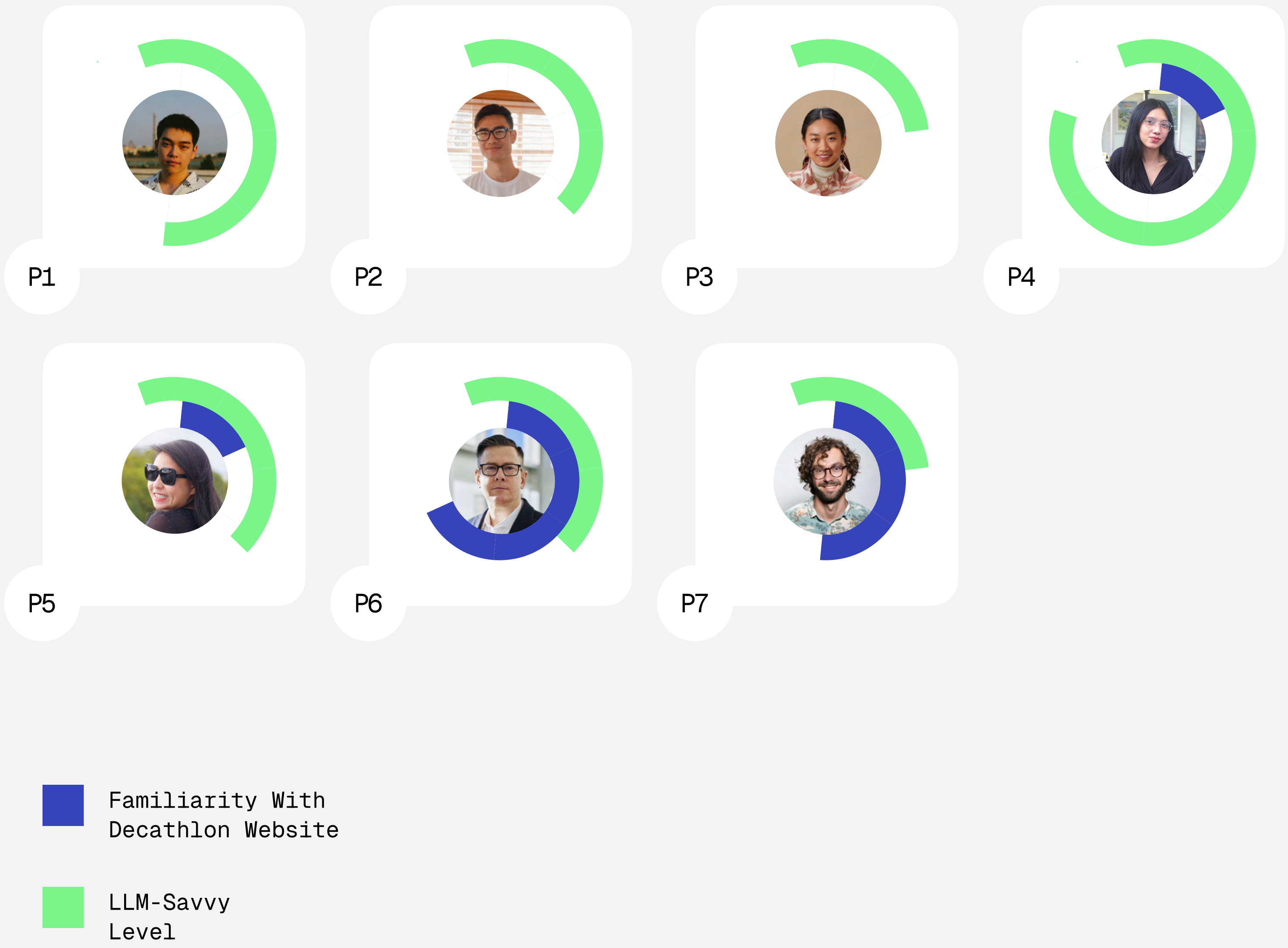
Data Collection

Given the exploratory nature of this research, for participants, I specifically sought individuals with both an interest in AI tools and regular online shopping experience to gather authentic feedback. Participants were recruited through two channels. The first was the WhatsApp group of TU Delft's Industrial Design Engineering Faculty, where I distributed an open call containing a concise introduction to the study's objectives and procedure. The second channel utilized Decathlon's Co-creation online platform, which provides researchers access to community members who have voluntarily registered their interest in taking participant in product research and development. These community participants typically engage out of genuine curiosity about Decathlon's behind-the-scenes development processes for both digital interfaces and physical products. They generally possessed more extensive experience navigating Decathlon's e-commerce environment compared to other participants. No monetary incentives were offered to any participants.

For both channels, the potential participants completed a screening survey indicating their gender, frequency of shopping on the Decathlon website, and familiarity with large language model technology. Due to time strains, in total, seven participants contributed to the study: five from the Industrial Design Engineering Faculty and two from the Decathlon Co-creation community. An overview of participant demographics and characteristics can be found on the following page.



Top: Visual representation of study participants. Note: Profile images shown are representative and do not depict actual study participants, used to maintain anonymity while illustrating the diverse perspectives that informed this research.



Based on results from the screening survey, all seven participants represent diversity across age groups, familiarity with AI technology, and experience with the Decathlon website. The familiarity with AI technology is based on their experience with ChatGPT-like AI systems, ranging from 'I have never used AI tools like ChatGPT or others before' to 'I have hands-on experience training or fine-tuning language models.' The experience with Decathlon website is based on their past shopping experience in the past 12 months, ranging from 'Never shopped at Decathlon website' to 'Multiple times per week.'

On the left is a visualization showing each participant's profile, with green blocks indicating their LLM-Savvy Level and blue blocks showing their Decathlon website familiarity.

According to Decathlon stakeholders, their primary user demographic is middle-aged males, which gives additional significance to Participant 6, who also ranks highest in familiarity with the Decathlon website. Given Decathlon's position as a provider of products and services for the mass market rather than niche segments, the analysis shouldn't exclusively prioritize feedback from specific participants. Nevertheless, it's valuable to observe a relatively normal distribution in the sample, indicating coverage across Decathlon's diverse consumer base.

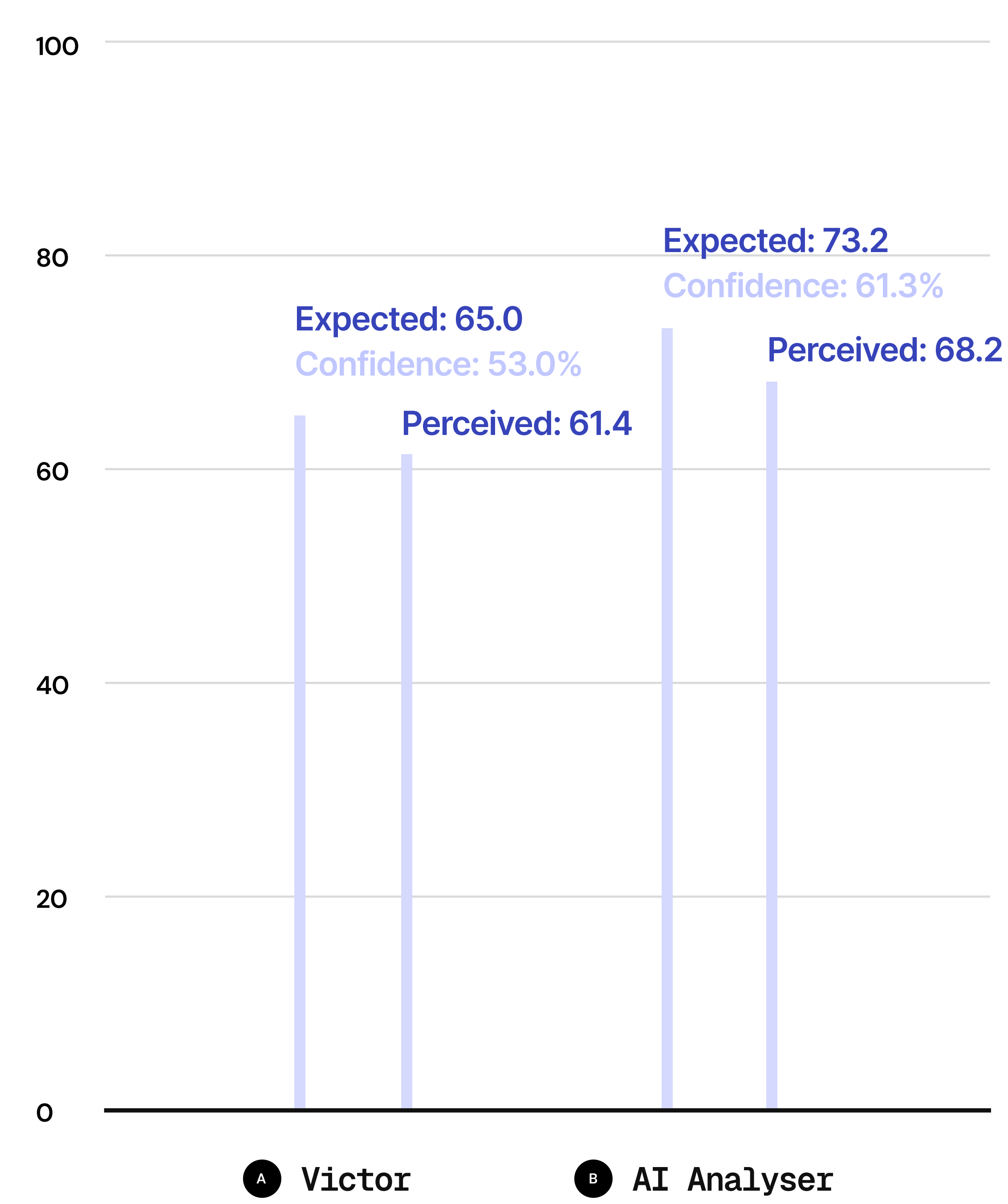
Data Analysis: Survey

Participants showed higher expectations and perceived functionality with AI Analyser, but neither exceeded expectations.

Contrary to assumptions, the 'AI Analyser' initially received a higher overall expectation score with a mean of 73.2 across the four dimensions. However, after users explored the prototype, the score dropped to 68.2, meaning it didn't meet expectations, creating a -5.0 gap. Meanwhile, for the chat interface "Victor", the initial expectation was lower than the 'AI Analyser' by 8.2 points. The post-experience score for the chat interface was also lower than the pre-experience perception, with a gap of -3.6 points. Although the gap is slightly less significant, the AI Analyser scored higher both in the pre- and post-measurements, compared to Victor, indicating a relatively better experience.

Regarding confidence in rating pre-experience perception, the results confirm our hypothesis that the AI analyzer's design enhances user understanding of its functionality. The AI Analyser received notably higher confidence scores compared to the Chatbot 'Victor'. Specifically, AI Analyser's confidence ratings fell between 'Neither agree nor disagree' and 'Somewhat agree' on the scale, while 'Victor' scored in the lower range between 'Somewhat disagree' and 'Neither agree nor disagree'. This difference in confidence ratings provides evidence that the visual design elements of the AI Analyser effectively communicate its intended purpose to users before interaction.

User Expectations and Experience Ratings for Both Prototypes



Examining the dimensions, AI Analyser exceeded expectations in interpreting intentions and providing accurate information.

For the first three dimensions, it also rated higher in post-measurement, indicating better performance, though it didn't demonstrate particular helpfulness in the specific task of comparing products. Notably, for the statement "It provided accurate information," the AI Analyzer showed an increase of 5.8 points compared to Victor's 1.4 points, demonstrating superiority in the perception of delivering accurate information. Another interesting finding is that for intention interpretation, both interfaces experienced an increase in scores, suggesting that both successfully captured users' intentions better than initially expected. On the other hand, both interfaces failed to make a good impression regarding "remembering context from earlier in our conversation," which is understandable given the design limitations. For the final dimension, the drop in scores may be attributed to feature limitations in the product comparison use case, especially for the AI Analyzer, which emphasizes its 'analysis' function.

User Expectations and Experience Ratings for Both Prototypes, Across 4 Dimensions

Regarding satisfaction, the AI Analyser was rated equal to or better than the chatbot across all the five dimensions.

Analysis was conducted by calculating mean scores for each prototype across five dimensions, using a 5-point Likert scale (1=strongly disagree to 5=strongly agree).

Most notably, for the statement "I am satisfied with how well it worked," the structured interface received ratings between 'Neither agree nor disagree' and 'Somewhat agree' (3-4 range), while the chatbot interface scored between 'Somewhat disagree' and 'Neither agree nor disagree' (2-3 range). For the "would use" and "would recommend" dimensions, differences between interfaces were minimal. The "found it helpful" dimension showed a relatively small but meaningful difference. Regarding the "found it to be annoying" dimension, neither prototype received strongly negative feedback, though a substantial difference was observed. The structured interface scored below 2, indicating users found it minimally annoying, while the chatbot interface received slightly higher ratings.

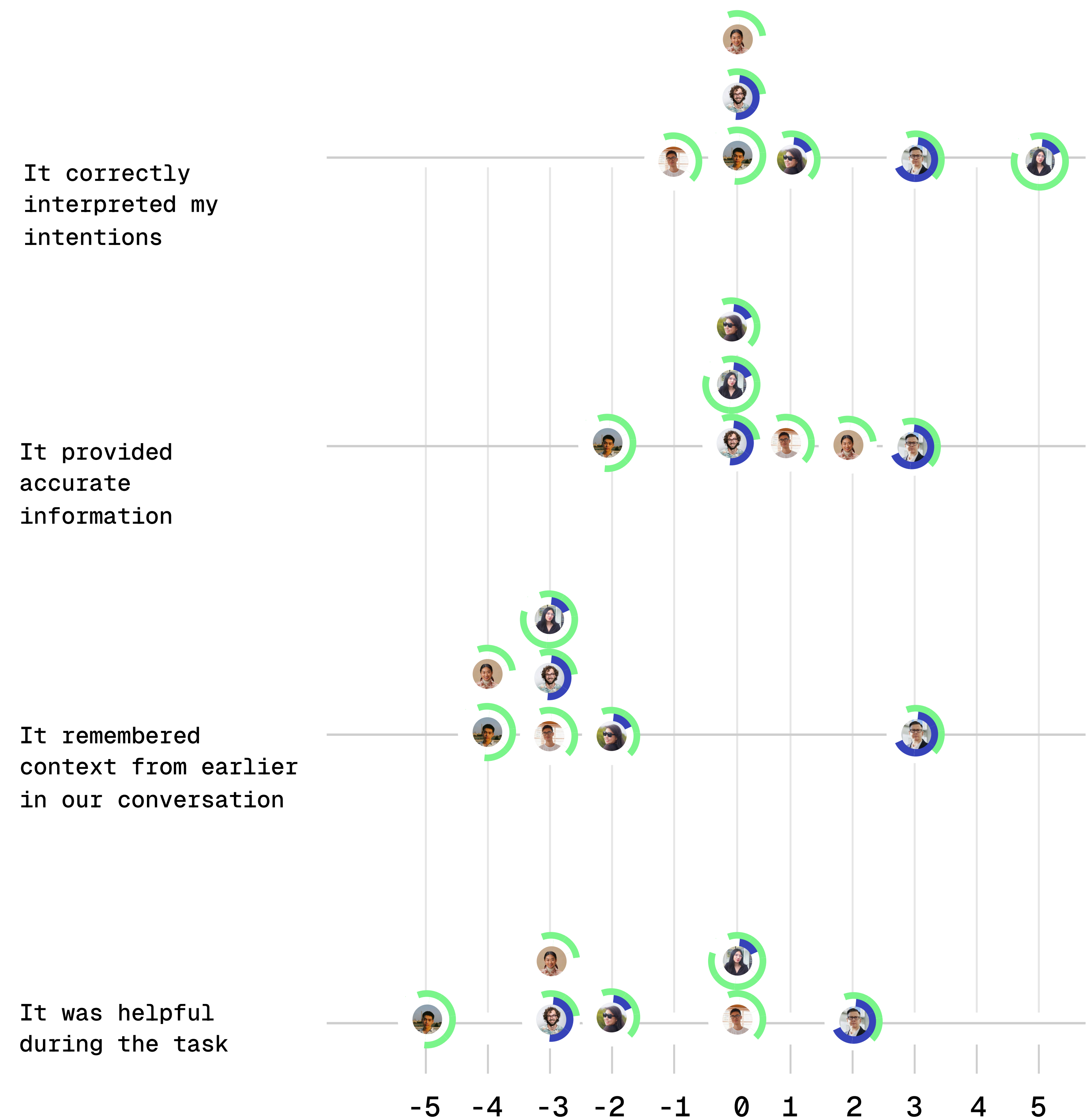
User Satisfaction Ratings Across Five Dimensions for Both Prototypes

Looking at individual participants, the experience varies.

Given the relatively small sample size, it was also interesting to examine individual participant data to identify potential outliers or other patterns. The visualisation on the right shows the results of analysing pre- and post-perception scores for the AI Analyser. When expectations are exceeded, the score would be higher than zero, thus the participant's identifier would fall on the right side of the graph.

Looking at the graph, notably Participant 6 shows consistently high expectations, which could be correlated to his familiarity with the Decathlon website flow, but this remains unclear due to the limited sample size.

Individual Distribution In Gap Between Perceived Experience And Expectation Across Dimensions



The survey is concluded by an additional comments area, where some of the participants have left their thoughts directly after taking the measurement questions.

I liked the honesty of the answers, especially based on existing real reviews

B AI Analyser

can designed better for comparing products

B AI Analyser

It would be great if it can allow me to compare 2 or even more than 2 products; also provide me with bullet points about features when I'm comparing (like Taobao)

B AI Analyser

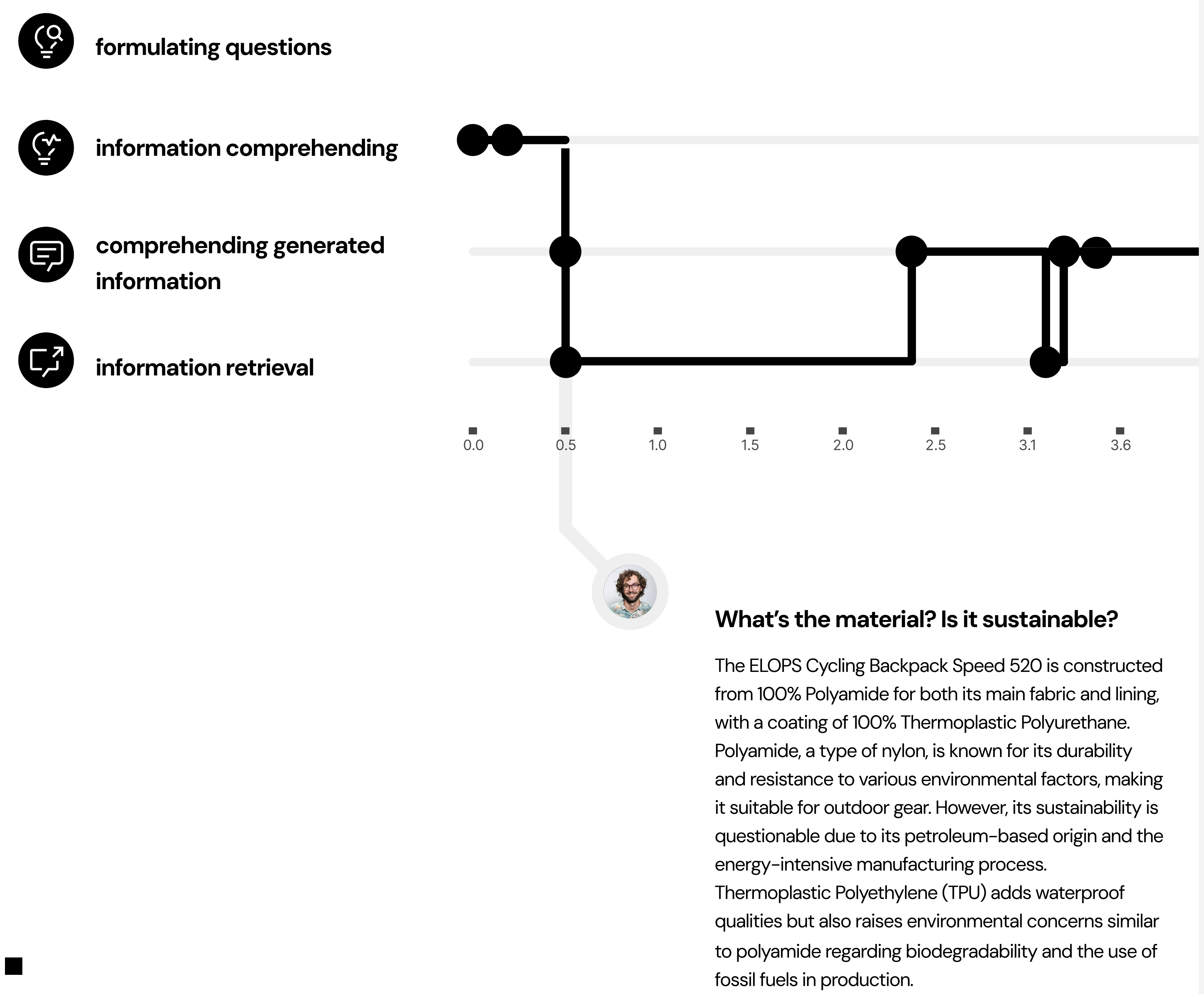
It would be great if the user interface could remember and repeat questions across different products

A Victor

hard to find the Ai

A Victor

Data Analysis: Activities



Besides the survey data, observations of user behavior were also recorded and provided insights into users' different navigation strategies. The recorded videos were manually processed by the researcher: each video were divided into segments representing distinct units of activity. For example, when a user was browsing the page rather than engaging with the AI component, these segments were categorized and logged into different action categories. Due to a researcher's mistake, participant 4's data were lost. Additionally, participant 3's data had to be excluded because the participant failed to follow the part of the study protocol. Therefore, both datasets were excluded from this analysis.

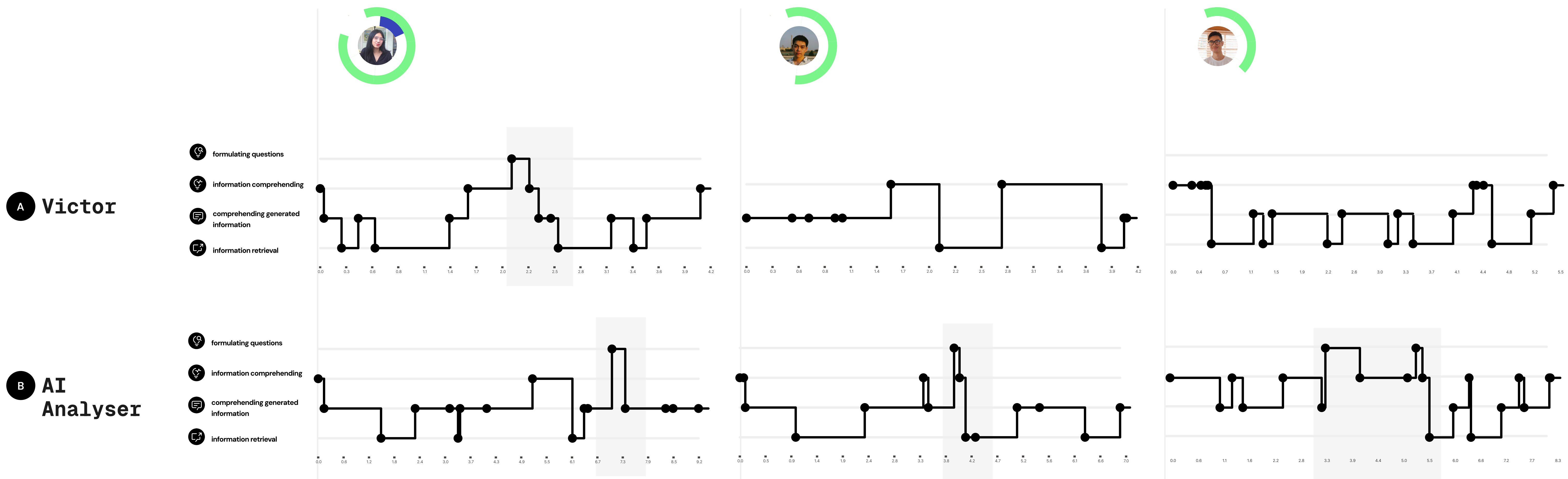
The action types and duration data were systematically recorded in CSV files. These files were then processed using a custom visualization script created by the researcher. The resulting web-based visualization tool allowed for efficient exploration and comparison of data across both prototypes and all participants. This visual representation of the interaction patterns significantly enhanced our ability to conduct the subsequent analysis.

On the right is an example of the activity analysis. On the left side of the map, there are four types of activities, namely 'formulating questions,' 'information comprehension,' 'comprehending generated information,' and 'information retrieval.' By distilling these different actions, the analysis reveals how participants navigate through the prototypes. Underneath the activity map, each participant's query and received output was also mapped. This data provides a more direct understanding of how users interact with LLMs through the interface.

Query formulation time - the friction that may created better input.

Three of the seven participants were observed to activate the input bar, then pause while typing, during which they looked through other information on the page. This observation suggests intentional formulation or cognitive processing, which raises questions about whether more input facilitation should have been provided.

This pattern was more frequently observed during AI Analyser interactions, suggesting either that the AI Analyser demands more thoughtful input, or that users wanted to carefully frame their queries to achieve expected results. The observed cognitive load may represent 'productive friction' - where the interface's complexity forces users to be more thoughtful and precise, leading to better responses.



Participants adapt their queries based on interface design- a subtlety that shapes different experiences.

Looking at how user inputs are formulated, it was observed that when interacting with Victor, some of the users would ask questions with 'Should I...?' but never used this phrasing for the AI Analyzer. This is interesting because higher-level questions like 'should I' presumably cost less mental resources by relying on the AI to incorporate reasoning and context about the user's situation. However, when it's unclear whether the system has contextual or reasoning abilities, the generated output doesn't necessarily provide helpful guidance to the user.

This observation supports the thesis argument that chat-based interfaces can set expectations misaligned with the system's actual functionality. However, it also advocates for future exploration in designing interfaces that better indicate system capabilities while maintaining intuitive interaction patterns. This would enable users to formulate questions naturally while understanding the system's boundaries, ultimately leading to more effective assistance.

The complete activity analysis maps can be found in the appendix.

A Victor

So, I shouldn't buy this right?

The ELOPS Cycling Backpack Speed 520 has received predominantly positive reviews, highlighting its waterproof capabilities, comfort, and visibility, with a high overall review score. However, there are concerns regarding durability and design, with some users experiencing wear and tear or finding the backpack's size and compartment design less than ideal. The decision to purchase should weigh these factors based on personal needs and preferences.

Should I buy it for my friend who lives in the Netherlands?

The ELOPS Cycling Backpack Speed 520 is highly recommended for its waterproof features, comfort, and visibility, making it an excellent choice for cycling in varied weather conditions, which is common in the Netherlands. The backpack's design includes reflective strips for enhanced night visibility, a waterproof seam, and a roll-top opening, ensuring belongings stay dry. Its ample storage space, including a laptop compartment and quick side access, adds to its functionality. However, some reviews mention concerns about durability and comfort under heavy loads, which are important considerations.

B AI Analyser

What do the latest reviews say about this product ?

The latest reviews of the ELOPS Cycling Backpack Speed 520 highlight its excellent waterproofing, comfort, and visibility, aligning with the product's advertised advantages. Reviewers appreciate its ample storage, including a laptop compartment and quick side access, which enhances its utility for cycling and commuting. However, there are mixed feelings about its design and durability, with some users noting issues with wear over time and suggesting improvements in compartment organization and strap design. Overall, the backpack is praised for its value for money, especially given its performance in wet conditions and its reflective features for safety.

Does this backpack fit the needs of those who live in the Netherlands? Why?

Given the Netherlands' rainy climate, the ELOPS Cycling Backpack Speed 520 seems well-suited for residents due to its 100% waterproof feature, as highlighted by multiple reviews praising its effectiveness in wet conditions. The backpack's reflective strips for visibility align with the safety needs of cyclists in the often overcast and dark weather. Additionally, the backpack's comfortable design and ample storage, including a laptop compartment, cater to the practical needs of daily commuters or cyclists in the Netherlands.

Data Analysis: Interview

For each participant, a transcript was generated during the user study and organized as docx files. These transcripts were thoroughly reviewed by the researchers and analyzed using the thematic analysis approach as outlined by Braun and Clarke (Braun & Clarke, 2022). This process involved six steps: familiarizing with the text, initial coding, generating initial themes, reviewing themes, defining and naming themes, and producing the report.

The analysis yielded valuable insights both directly related to the research questions and beyond their initial scope, revealing unexpected topics worthy of exploration. This discussion will first address the supportive evidence for the hypotheses, then examine the emerging patterns found.

KEY INSIGHT 01

The demo effectively addresses small but time-consuming validation needs

KEY INSIGHT 02

Transparency-focused design improves user experience quality, despite low feature utilization

KEY INSIGHT 03

Structured interface could potentially reduce bias perception and increase user confidence

KEY INSIGHT 01

Let's start from the first hypothesis -

Hypothesis 1: Integrating contextual LLM capabilities directly into product pages will reduce cognitive load and increase information discovery efficiency compared to traditional browsing approach.

Although the quantitative analysis show that for AI Analyser the helpfulness has a mean of 60 out of 100, which is relatively low, the research still found strong support for this hypothesis through multiple user comments, especially supported by the participant 6 when reflecting on how AI Analyser could be helpful in his journey on finding right shoes:

> **" the first thing I do is... especially the ones that have a lot of reviews, I go down and I try to see what are the negative ones. So, if there is a way that an AI tool can actually filter the reviews and give me the reviews that are more relevant to my question, this could help a lot instead of me trying to scroll through 1,000 reviews"**

The user also mentioned the problem of multilingual content, which can be tedious to translate everytime

> **" this especially for products that are sold for years, you can have thousands of reviews. So in different languages because Decathlon is also selling in different countries. So very often you try to filter reviews and you get one in Spanish, five in French and one in Dutch."**

Moreover, the participant described the time-consuming process of researching technical shoe specifications:

> **" when I to try find the shoes of Decathlon and usually you spend quite some time. what is the drop of the shoe? Is it 6 mm or 8 mm? or for example you have two forms one is called M form and the other which is the traditional old-fashioned form and another called V form which is the most advanced one then I spent quite some time when there is a new model, I will try to see in the description of the product, so if I could throw these questions to AI it would save me time immediately "**

The evidence suggests that shoppers who frequently make purchases and care about specific model features, design details, and technical specifications would particularly benefit from LLM as an automation tool. Besides the frictions in user experience and the technical limitations of the prototype used in this study, by enabling direct queries about product attributes, an LLM integration could significantly reduce the time spent searching for information within various sections of product pages.

KEY INSIGHT 02

Hypothesis 2: Structured, non-conversational patterns like the FAQ approach will lead to higher user satisfaction for product-specific inquiries.

Besides the positive quantitative results, user comments supported this hypothesis, with participants specifically praising the design elements of the AI Analyser interface:

> **" I don't know why, for some reason, it feels like it's less of an effort, maybe because it's actually just one click "**

> **"but honestly I like this. I like this interface. I like the history section. I can see a lot of previous chats"**

> **"Oh i like the history, and i like the tags"**

These comments reflect a consensus appreciating design elements that address unexpressed user needs - particularly the ability to view conversation history and explore functionalities with minimal effort. However, it remains unclear how these functions actually supported users in their decision-making process, since most participants didn't discover the history panel or interact with it.

KEY INSIGHT 03

Hypothesis 3: More constrained interaction patterns will create clearer mental models for users, resulting in fewer misunderstandings about system capabilities and more accurate predictions of system behaviour.

This hypothesis is only partially supported because the name of the AI Analyser has caused confusion, leading users to believe that it has more sophisticated features:

> **"I tried the AI analyzer first and I found that it's not super helpful because it cannot help me compare between different products."**

Despite shared confusion about the knowledge boundaries of both LLM prototypes (Theme 3), users demonstrated different mental models for each system. One participant articulated this nuanced difference in terms of perceived value: Victor was associated more with decision-making assistance, while AI Analyzer was perceived as a tool for accessing product details.

> **"So, the first one is a straight up ask a question. Some questions are selected for you interaction to decide on your product purchase, which I don't find to be appealing for me. whereas the other one is just another way of presenting the product details. and uses questions which I don't know I kind of want to click on it"**

The comparison from this user reveals how interface design influences expectations about system functionality. The more constrained AI Analyzer interface appears to suggest better autonomy, whereas Victor creates a sense of bias. Although both systems use the same suggested questions, the context in which the AI is presented creates different user perceptions.

Furthermore, the interviews revealed additional insights into the landscape of human-LLM interactions, which I found refreshingly grounded in actual user experiences. I have synthesised these into thematic patterns.

KEY INSIGHT 04

The name of “AI” has been confusing users.

KEY INSIGHT 05

The gap between LLM’s promise of 'knowing everything' and users' need to understand its actual limits creates fundamental usability challenges.

KEY INSIGHT 06

Consider both distributed and global LLM workflows.

Theme 1:
Heterogeneous user experiences and definitional ambiguity in broadly-categorised AI products in the market

People commonly make sense of new technology by drawing on their past experiences, which we observed in our experiments where users compared the LLM prototypes to technologies they were familiar with. However, we found no consensus on how these prototypes related to current state-of-the-art LLM applications. For instance, Participant 1 felt more comfortable with the AI Analyzer because of its similarity to ChatGPT:

> **“ This looks more like ChatGPT so I feel more confident with this one”**

Conversely, Participant 2 expected Victor to be more intelligent for the same reason:

> **“Victor is very obvious like it is a AI agent... it looks like a chatbot, so I expect it to be more intelligent, smarter.”**

Within the study's constraints, it was difficult to pinpoint exactly what features participants associated with the concept of an AI agent. What became clear, though, is that no "prototypical" AI agent interface exists yet in users' minds. Part of this confusion stems from the blurred distinction between rule-based chatbots and LLM-based chatbots. Participant 2 illustrated his previous experience with rule-based chatbots to explain his expectations for Victor:

> **“The usual average experience I had with chatbot is like it has 100 holes, and it is trying to figure out which hole I fit in.”**

Other participants similarly shared stories of their encounters with customer service chatbots when reflecting on their previous interactions with AI:

> **“I don't think I used... any AI (for shopping). Only when I have specific questions for logistic or delivery, I will ask the customer service.”**

> **“It reminds me of chatbot that always are preventing you from talking to a real person...It's a bit old-fashioned.”**

The experience with rule-base chatbot has lowered their expectations of the product when it’s designed with a similar chatbot interface—a finding supported by the previous quantitative analysis.

The confusion is further attributed to the widespread implementation of general-purpose Large Language Models across numerous common applications.

KEY INSIGHT 04

Participant 7 articulated experiencing cognitive fatigue from encountering LLMs in multiple digital contexts despite their perceived limited utility, which significantly influenced his evaluation of Victor due to apparent categorisation within the same technological classification.

> **“because I'm a little bit tired of seeing these LLMs everywhere and I don't see their value....The Victor button reminds me of the AI on WhatsApp, and it really annoyed me. so now every time I see that kind of blue bluish I'm like, oh this is AI...”**

These user comments particularly excited me as they revealed such divergent perspectives on the single two letters 'A' and 'I,' which have been repeatedly used across countless contexts for vastly different purposes. Yet we still haven't found effective ways for customer-facing products to differentiate these LLM implementations in ways that help users form clearer mental models.

On a self-reflection note, perhaps it would be more accurate to describe my prototype as a 'text transformer system integrated with user reviews and product specifications' rather than simply calling it 'AI Analyser'. This wouldn't come with three sparkling stars next to it, but it would reduce the confusion.

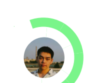
Theme 2:
The lengthy text prevents the users from comprehending the summarisation

As participants noted,

> **" I don't really want to see such a big paragraph like explain things in such a detail and in a in a paragraph I want to see bullet points"**

> **" .... actually, to be honest, when I first saw it I was like, this is so long and I skipped it, I didn't notice that is like the text that come from reviews. "**

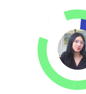
Rather than simply reducing length, we should consider the relationship between user intent and information presentation. When someone approaches content with careful examination in mind, comprehensive text serves their purpose. However, when seeking quick insights or comparisons, that same volume becomes a barrier. This suggests our challenge isn't just about length but about matching information density to diverse user needs and contexts. But at least in this use case, what we learned from users is that we should further design and evaluate different formats rather than defaulting to text blocks. Besides implementing bullet points, we might experiment with visual elements like the emojis one participant specifically mentioned.



> **" ...I thought it's good at organizing stuff to certain bullet points....because that's what my experience have been with ChatGPT. There is usually some bullet points and a final conclusion....Probably with some emojis that actually help you with recognizing. "**

Previous studies also demonstrate that emojis significantly enhance comprehension in digital environments (Hancock, Patrick M. 2023). This provides empirical support for the participant suggestions and offers a promising direction for improving engagement through visual elements alongside textual content. This is also an empirical support for our previous discussion on human memory limitation in tasks like shopping.

Besides, we should consider not only alternative ways of constructing the synthesised text, but also how to handle special content elements such as product names, as one participant encountered friction in memorising the product names in order to compare between them.



> **"It is hard to compare because I have to type the name of the products and they are very long"**

For highlighting special elements like product names, conventional approaches such as bullet points or emojis may prove inadequate. This insight points toward the need for deeper exploration into more sophisticated and intuitive highlighting techniques. Future research should investigate methods that elegantly emphasize critical information while maintaining textual coherence and professional presentation.

Theme 3:
Users' different interpretations of knowledge boundaries reveal the fundamental limitation of LLM

KEY INSIGHT 05

One participant emphasised that text should be concise when the system has limited information:

> **"I don't need it to say the things after the "I don't have the information", the words after are useless"**

Interestingly, another participant interpreted similar limitation acknowledgments quite differently. When encountering an LLM response stating limited information about sustainability, she perceived potential bias towards business purposes:

> **"I feel like the answer is biased, because when i asked if it is sustainable, it says that we do not have the information, (which means it is not sustainable)"**

This captured an interesting difference between users' mental models towards the system, showing that to achieve more nuanced expectation setting, the system might need to better explain not just what it doesn't know, but how it knows what it knows. Additionally, when the system lacks information, it should indirectly help shape an aligned mental model for users to prevent misinterpretations about intentional information withholding or bias. Furthermore, some users experienced confusion when asking for direct product comparisons, as they assumed the system had global understanding of all products, even for the AI analyzer which was more integrated into the product page:

> **"I feel like I only need to talk on one page, because it would know about other products too "**

This reflects both the technical limitation of the design and a gap in communicating knowledge boundaries to users. From the users' naturally formed mental model, we observed that there was insufficient guidance in clarifying the boundaries of the system's knowledge.

Another observations is without directly indicating the capability boundaries, users spend time on "testing" the boundaries themselves

> **"So, the thing is that usually with these LLMs, I like to have fun with it. So, I'm kind of…testing the boundaries of what it's going to give me. "**

These comments from users reveal a new dimension to consider when designing LLM-based systems. It is nearly impossible for end-users to understand how much data a model like GPT has been trained on, which creates a huge gap between the promise of 'knowing everything' and the reality that users need to understand what the model actually knows in order to use the tool effectively.

These findings reveal a design challenge: effectively communicating knowledge boundaries in ways that align with diverse user expectations. An effective interface must acknowledge information gaps if there is, and also provide appropriate context and facilitation for these limitations.

Theme 4: Trust boundaries of LLM technologies

Through the semantic analysis, I discovered that some of the participants established trust boundaries when using LLM technologies for product research tasks. They consistently demonstrated trust in AI for data summarisation capabilities, forming a baseline of expected functionality.

> **“I expect it to show the specs, then I would feel like it’s very professional, and I’m not “scammed””**

> **“But if I ask AI for factual information like material I would (trust it)**

> **“(I like to see) for example, this is the text from AI and this is a fact from review and then users can decide by themselves. ”**

> **“the first thing I do is go down and I try to see what are the negative ones (reviews). So, if there is a way that an AI tool can actually filter the reviews and give me the reviews that are more relevant to my question, this could help a lot instead of me trying to scroll through 1,000. ”**

These findings suggest that in this specific context, the users trusted the system primarily as an information processing tool rather than a decision-making authority. One participant shared her personal habits with online shopping, where she often shares options with friends for supportive judgment. When asked "Would you trust AI if AI does this recommendation for you like your friend does?" the participant responded:

> **“I don't think so. Because yeah, it's not someone that I spend most of the time with.”**

This raises a more philosophical question: if the LLM system could collect users' contextual information such as personal values, life histories, and preferences, would we trust a machine to make decisions for us? The discussion could go on and on, but here, with a surface-level understanding of the participants' comments, we can conclude users can be skeptical of the recommendations generated from a LLM-based system.

Another insight gained from the user study was that trust can also be influenced heavily by specific instances of experience. For instance, a participant share that:

> **“....when I use the AI assistant, for example, I opened the second page and it recommends that one and then I switch to another page and it recommends(the other one)... since then I don't trust it anymore, you know?”**

This observation highlights a critical tension in AI design: the balance between contextual adaptation and perceived consistency. While AI systems may be designed to provide tailored recommendations based on different products' unique attributes, users can interpret shifting recommendations as inconsistency or unreliability.

Another interesting finding is that participants found the term 'AI Analyzer' more trustworthy as a name because it suggests more fact-based and data-driven functionality. However, this naming also created higher expectations among users regarding the system's analytical capabilities.

> **“I have higher expectation on the AI analyzer because it's like more rigid, so I feel it's more convincing.”**

> **“maybe AI analyzer has more powerful database or maybe it can tell you like For example, 70% of users choose this other than the other one.”**

Lastly, one participant reported that he appreciated the "honesty" of the responses (see the previous “other comment” page) as the system clearly stated the negative reviews when prompted to do so. This indicates the unbiased information presentation can be critical in building trust with consumers.

To conclude, users' trust in LLM-based systems is a topic that falls outside this thesis's scope. From these dispersed comments from users, we can only form a vague understanding of how users guess what AI can do, which links back to the first theme: vaguely defined AI products.

Theme 5: Consideration of user flow and separately integrate AI related components

Another theme emerging from the user study was that participants showed preferences for both prototypes based on how each integrated into their overall workflow. The effectiveness of each approach varied depending on the user's immediate focus and task context. For example, the floating global position proved particularly valuable when users needed to compare multiple products simultaneously.

> **“ I can just anytime click on the this AI thing. Do this thing and check but here I can't do that. I have to like Scroll down the whole thing again so it was easier for me to find the AI in this than this.”**

> **“Maybe the AI should be on the previous (catalogue) page so that I can just directly search for it (product names)”**

These insights suggest that when designing LLM-based features, we should carefully consider where the functionality is located, and thoroughly examine what information users need to bring into their interaction flow. Additionally, we should evaluate how it might be convenient for users to have contextual information readily available alongside the LLM-based interface.

For AI Analyser, there were also participants who preferred this approach over the “global” position, particularly appreciating the open-and-close interaction pattern, which they felt provided clearer boundaries for the AI functionality.

> **“I really like the second one because I only had to use my mouse and click on a plus button that provided answers even though the answers could be a little more synthetic, a little more clear. I feel like that's a medium that I respond to much more. ”**

> **“So the fact that I had the questions down the menu and I could click and I could see that if it is not relevant for me, it just expands, I find it less intrusive....I like the first one, because it was a little bit more integrated”**

We could see that participants' responses focused on the actual evaluation of the product, which became more acceptable when the the LLM feature was less intrusive, rather than having a fixed position that remains constantly visible. Users appreciated how easy it was to close and open the integrated LLM feature.

This posts a distinctly different preference from those who preferred the “global” approach. The users who described the preference for chat interface valued being able to engage with AI at any point in their workflow or having access to a broader context that helped them formulate their questions.

Additionally, participants noted that the less intrusive design approach of AI Analyzer facilitated easier strategy shifts during the shopping process. Rather than becoming locked into a conversational flow, users appreciated having a tool that provided supportive data while allowing them to maintain control of their shopping journey.

KEY INSIGHT 06

> **“If it helps a bit, it's positive. If it does not help, you can always continue your shopping in the old way. That's how I see this. ”**

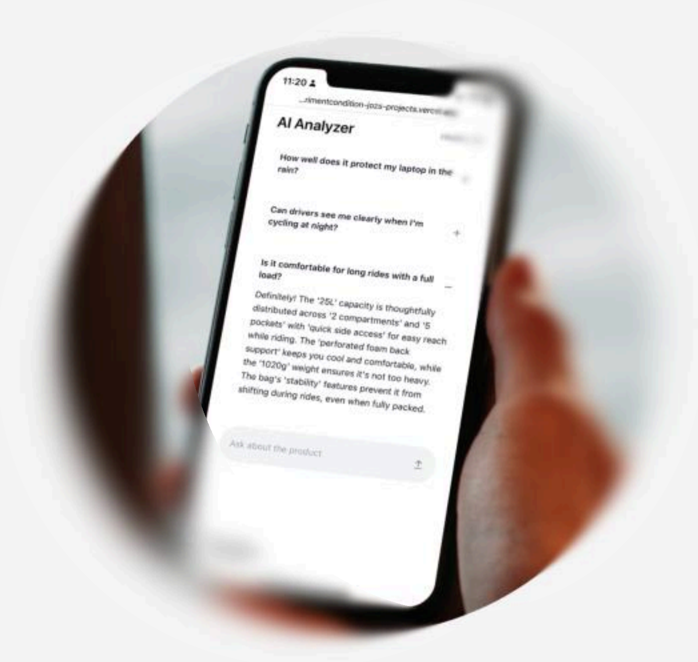
To conclude, what we could learn from these user comments is to re-think the design on a system level, considering the position of LLM feature in a complete user journey. This approach would allow us to better integrate LLM-related activities—such as formulating prompts, asking questions, and reviewing responses—into different parts of the user journey, rather than designing LLM product, enabling the right action at the right moment.

Takeaway

Static Webpage

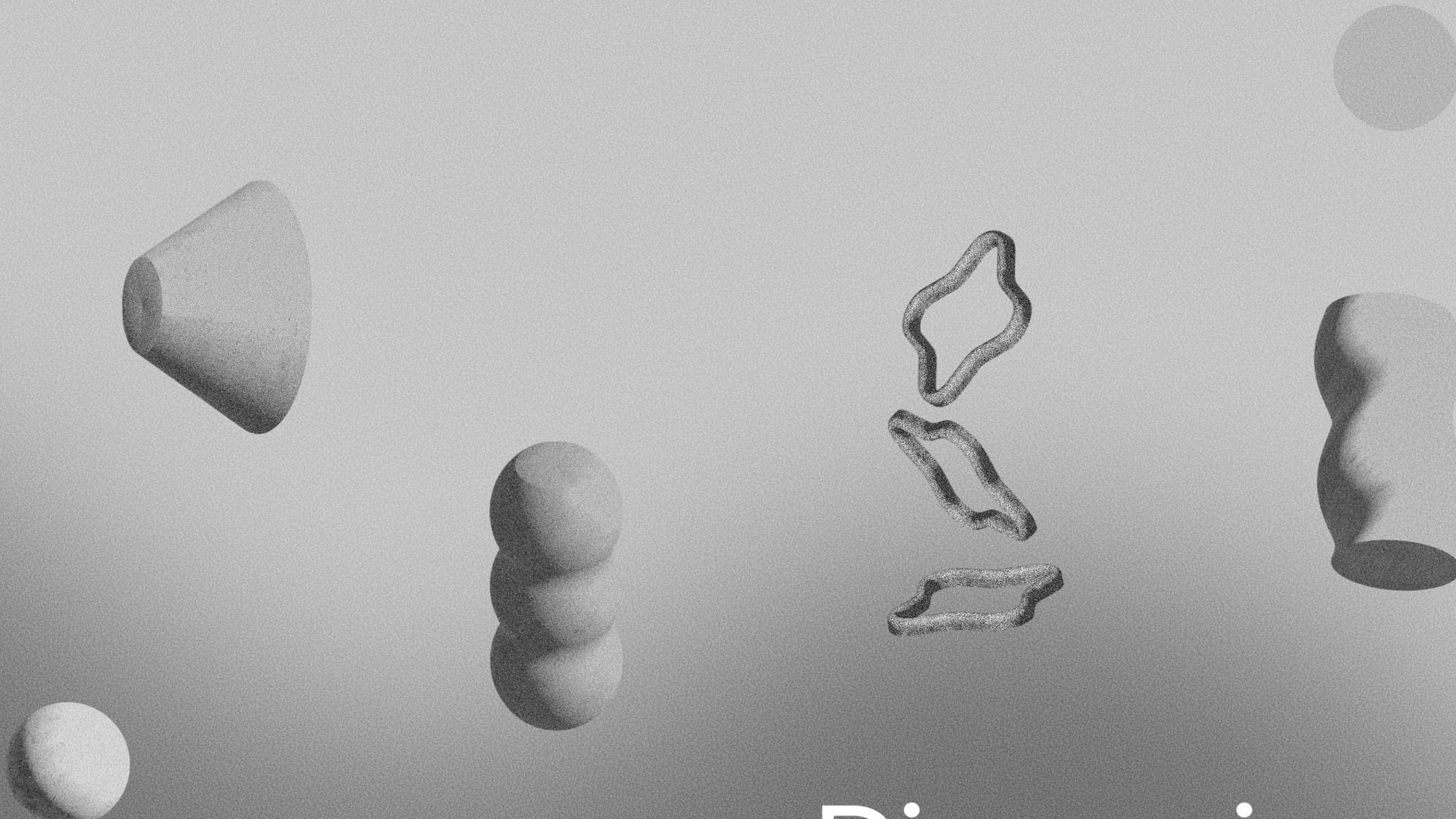
Does it solve the problem?

Yes



Chat-based Interface

Although the structured interface showed higher satisfaction and usability scores in the comparative study, the user study also pushed me to think beyond these results and imagine better interface possibilities.



Discussion

5.1

Design Guideline

Integrate research insights into actionable suggestions for future design

5.2

Roadmap

A handy roadmap for designing LLM-based systems in a human-centred approach

Actionable insights.

In this chapter, we'll elevate the insights we've gathered by exploring organisational perspectives through expert interviews. We'll then synthesise these findings into a practical design guide that transforms user insights into actionable design recommendations. Synthesising the insights from the study, it's valuable to develop an overview of key considerations when designing such functionality. Design guidelines provide an effective method for this approach.

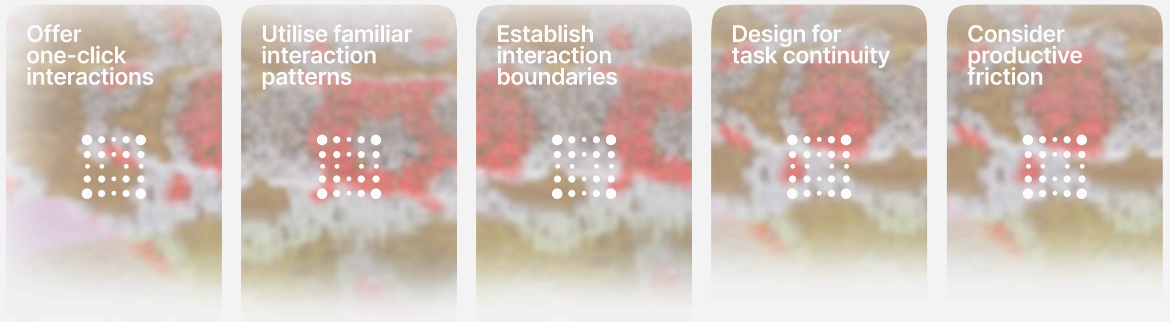
These suggestions emerged into four identical layers presented on the right.

1 Content

How might we make the LLM response more comprehensive?

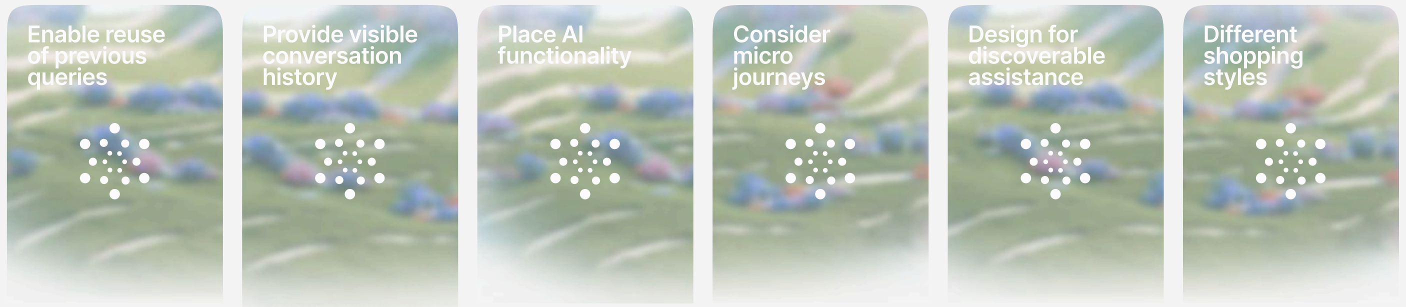
2 Interaction

How might we facilitate better understanding of the feature?



3 User Journey

How might we better integrate the feature into the user journey?



4 Positioning

How might we strategise the”feeling” of LLM?

Content:

How might we make the LLM response more comprehensive?

Despite hearing from users about how visual indicators could be helpful in LLM interactions, solid information architecture design is fundamentally grounded in understanding what makes information more comprehensible, drawing on theories such as Gestalt principles.

With the integration of LLM technologies, it becomes interesting to identify interdisciplinary design methods that would allow the system to automate basic design techniques such as clustering and contrast.

Furthermore, the depth of information needs specific examination of the context - for instance, in scenarios such as no results or quick validation needs, it's worth exploring design strategies for making responses more compact.

Add visual indicators

Add visual indicators like emojis to help users quickly distinguish between different types of information

Keep text comprehensible

Keep text concise for quick scanning, with the option to expand for details

Highlight special elements

Highlight product names and technical terms for easier mental processing and comparison

Tailor information depth

Tailor information depth based on product complexity and purchase frequency

Interaction:

How might we facilitate better understanding of the feature?

On the interaction level, universal suggestions are tailored first to proper mental model building - when users send requests that are out of scope, reminders could better shape subsequent interaction rounds.

Additionally, the user study showed that FAQ-like interfaces facilitate better usability when users are in the context of product pages and are familiar with these interaction patterns, which could be extended to other scenarios.

I also discovered design choices that make sense in different contexts. For instance, the balance between "one-click-away" interactions and "productive friction" can provide benefits in different ways.

Establish interaction boundaries

Establish clear interaction boundaries through concise, friendly responses when questions fall outside the system's scope or redirecting to appropriate channels for out-of-scope requests

Design for task continuity

Design for task continuity by considering what users do before and after AI interactions

Utilise familiar interaction patterns

Maintain interaction consistency with familiar browsing patterns to reduce learning curve

Offer one-click interactions

Offer one-click interactions whenever possible to reduce perceived effort and increase engagement

Consider productive friction

Allow strategic cognitive effort that leads to better outcomes.

User Journey:

How might we better integrate the feature into the user journey?

Thinking on the user journey level can also generate ideas that place users at the center of LLM-based systems. This allows moving attention from the LLM itself to a bigger picture where users are already engaged in certain flows of actions.

In this context, the strategy becomes blending AI into existing flows - when should it show up? Where should it show up? When should it disappear and reappear? Do people have varying preferences?

Consider micro journeys

Consider "micro journeys" by integrating LLM access points at various stages of the shopping experience

Design for discoverable assistance

Design for discoverable but non-intrusive assistance that complements rather than disrupts established workflows

Place AI functionality

Place AI functionality where it's most contextually relevant (product catalog, comparison views, detail pages)

Provide visible conversation history

Provide visible conversation history to help users track their information journey

Different shopping styles

Respect different shopping styles by offering multiple ways to access AI assistance and making it easy to switch between them

Positioning

How might we strategise the“feeling” of LLM?

The last layer appears to be an emerging area for consideration with LLM technology - these suggestions only capture a small portion of the landscape of regulating LLM outputs. However, in this specific context, users have shown that it's important to maintain consistency and neutrality in LLM behaviour, which forms the foundation of trust on the system.

More fundamentally, the research observed the confusion LLMs create regarding their knowledge boundaries. Although transparency issues can be partially mitigated through active interactive responses, it's worth exploring how to fundamentally address this problem to ensure users are aware of technical possibilities and boundaries.

Lastly, an issue that might be temporary but poses high value is avoiding designs for similar applications that would trigger negative associations and get an immediate negative response. This could be addressed through market trend analysis, technology scouting, and risk mitigation strategies.

Maintain consistency

Maintain consistency in recommendations to build trust

Present balanced information

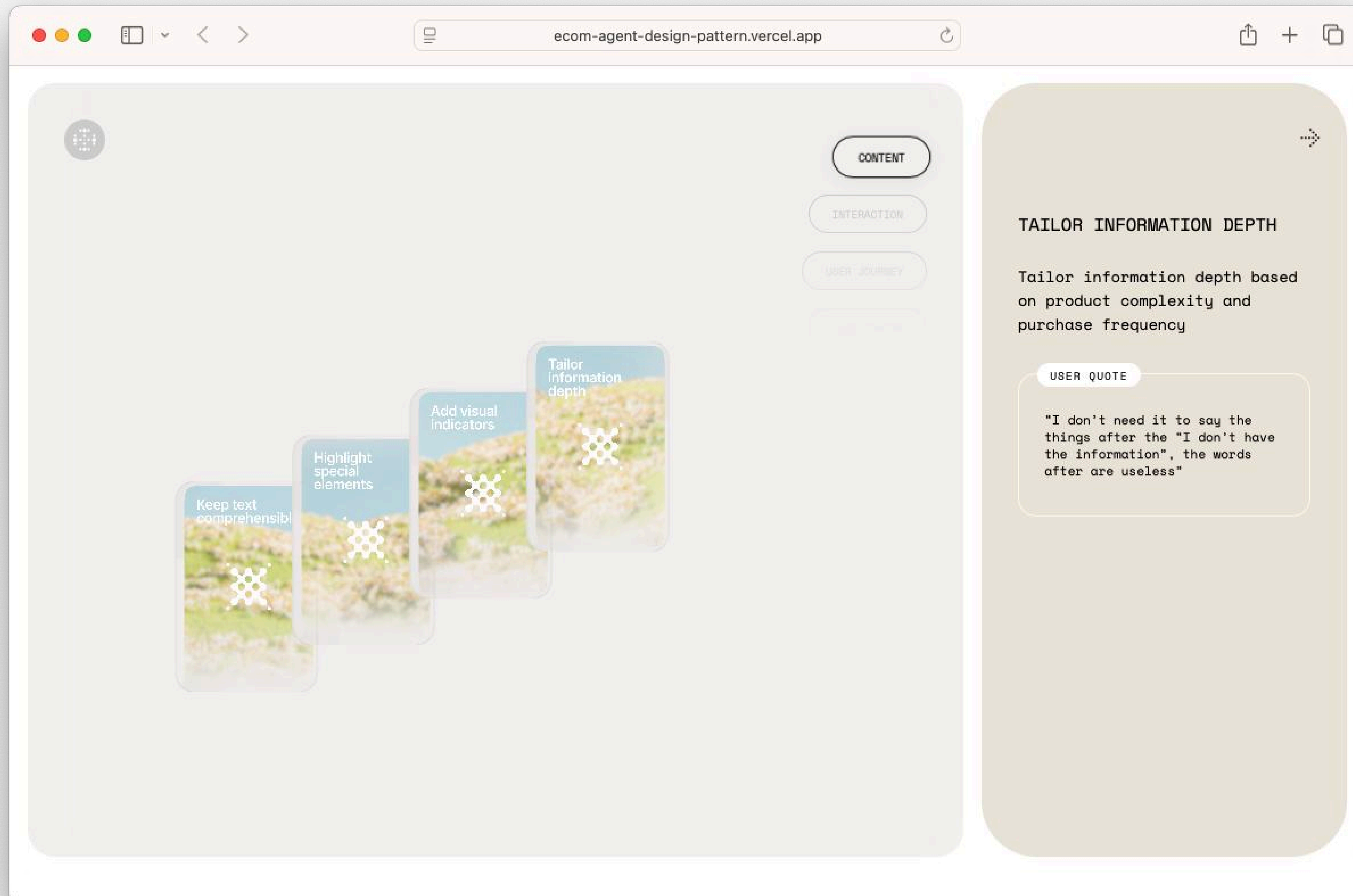
Present balanced information including both positive and negative aspects of products

Proactively explain boundaries

Proactively explain knowledge boundaries of the LLM, provide clear, upfront explanations of what data the system has access to

Avoid triggering negative associations

Avoid triggering negative associations through researching common pain-points with other AI tools and deliberately design differently



The design suggestions were also synthesised into an interactive artefact, the artefact can be accessed through: https://ecom-agent-design-pattern.vercel.app

A user-centric roadmap for LLM system exploration

The user study revealed how interface variations quietly stimulated participants to generate different queries, form different views on LLM output, and even develop attitudes toward the brand such as perceptions of honesty or market orientation. This supported the hypothesis that interface design significantly contributes to user satisfaction factors.

For non-LLM developers, the design guidelines and roadmap open up opportunities for innovative interface design patterns that reconcile limitations inherent in LLM technologies, such as linear interaction constraints, unknown knowledge boundaries. For instance, in recent practice, exploration with teams has focused on combining graphical user interfaces with linear chat interfaces within holistic user flows, allowing the graphical interface to adapt to interactions happening in the natural language interface. Through clear definition of the user needs to be met, it becomes natural to further define which interface would be optimal at certain moments. With this basic structure, we can then fine-tune interaction details such as creating experiences for setting mental models of knowledge boundaries, improving the comprehensiveness of generated content, and establishing a consistent tone of voice.

This thought process can be useful when creating unique experiences that won't elicit a **'some kind of AI again?'** response from users. The following page sketches out a roadmap that guides exploration of designing LLM-based systems in a user-centric way.

Step1: Idea

You may say that "we should build a chatbot."

Do we know what problems we want to solve?

Not sure

That's okay, let's play with it. It's helpful to understand what LLM could do first, and connect it with existing user needs.

Yes, for example, we want to use language model to answer questions about the product to save users' time in browsing the pages.

Are there more user-friendly design patterns than chat for this use case? If yes, can we integrate them into a hybrid pattern to make it more familiar to the user?

Step2: Design

When designing the LLM part...

How might we make the LLM response more comprehensive? (e.g., prompt engineering techniques)

How might we facilitate better understanding of the feature? (e.g., establish boundaries, contextualization)

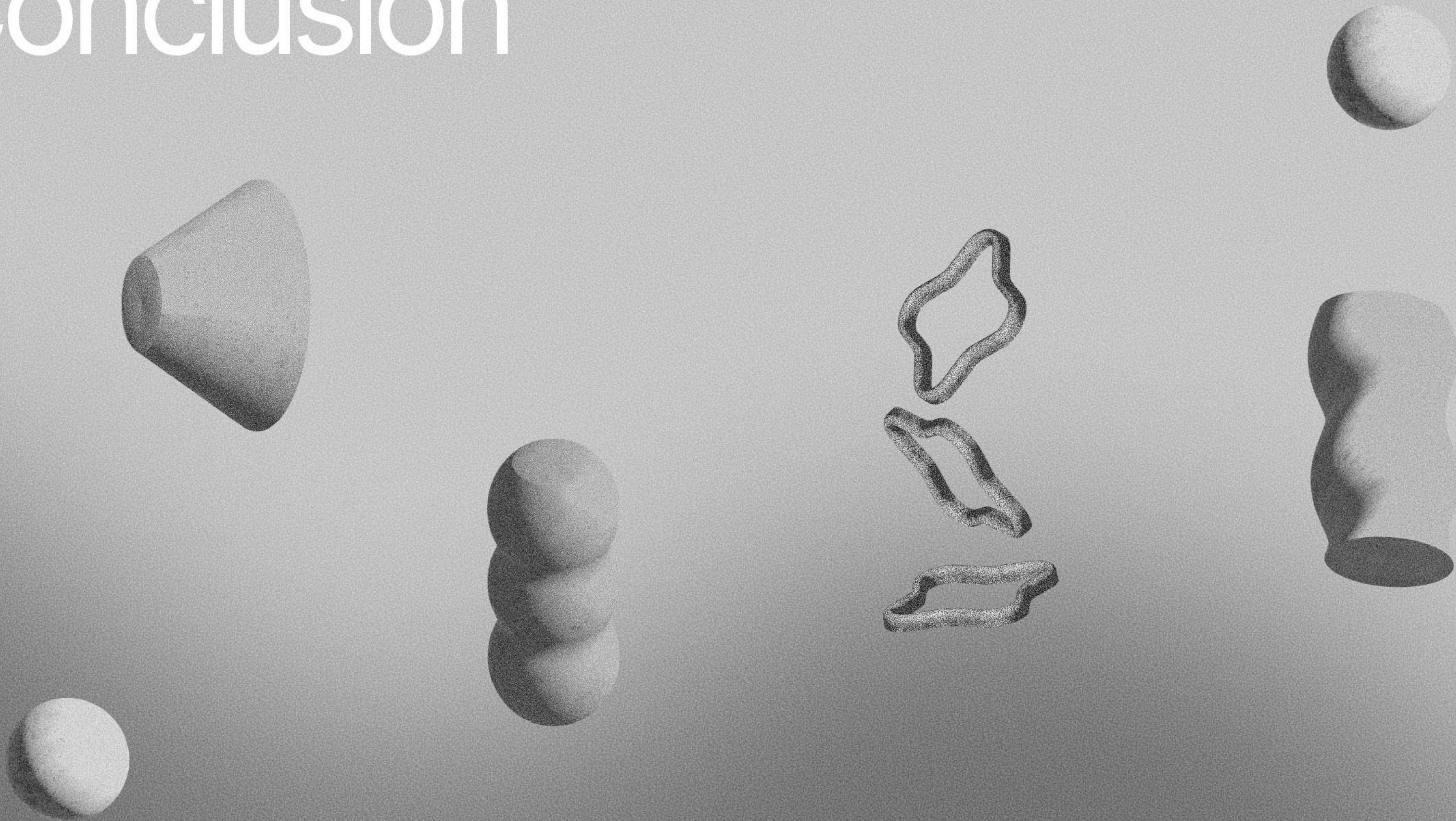
How might we better integrate the feature into the user journey? (e.g., any additional friction?)

How might we maintain a consistent "feeling" through the interactions? (e.g., honesty)

Step3: Test

Don't forget to test your product (in your way)!

Conclusion



6.1

Overview

An overview of the research

6.2

Limitations & Future Directions

Address limitations identified in this research and suggest for future explorations

6.3

Reflection

Reflect on the project on a more personal level

Overview of the research

This research aimed to explore the possibilities of designing LLM-based systems in the context of digital commerce platforms that would help users navigate purchase options, with the additional intention of exploring better human-LLM interaction design patterns. The exploration included both theoretical investigation and empirical design experiments with user studies. In the end, design suggestions were made that are grounded in theories and empirical findings.

I delved into LLMs' practical potential within online shopping environments, where capabilities like taxonomy categorization, information extraction, and contextual sensitivity have shown potential in helping users navigate overwhelming product information landscapes.

Through systematic investigation, I uncovered fundamental issues: misalignments between user mental models and system capabilities, tension between linear information presentation in chat-based interfaces versus humans' innate preference for spatial information processing; and on the potential of LLMs - potential cognitive load increases, deeper understanding of what actually helps in consumer decision-making.

With these insight - understanding both potentials and limitations - I explored alternative interfaces beyond chatbots through human-centred design methodologies. This foundation informed my rapid prototyping of four distinct LLM-integrated concept.

The final implementation, AI Analyser, integrated user reviews and product information into a FAQ-like components that allow the users to request for detailed product information. AI Analyser demonstrated three key functionalities: answering product-specific questions using review and product databases, preserving past queries, and providing a history panel revealing system knowledge of user queries and summarised preferences. To understand how this approach compares with the state-of-art chat-based interface, a chat-based interface was prepared before the user study, that integrates the same technical architecture.

With both interfaces ready, I designed a mixed-methods user study combining task-based randomized A/B testing with semi-structured interviews, ultimately engaging seven participants.

The collected data revealed that the structured interface quantitatively demonstrated improved overall experience, and enhanced perception of both information accuracy and intention interpretation.

The analysis of participants' behaviour reveals the possibility of productive friction increasing the usability of the system and how different interfaces prompted for different ways of query formulation.

The interview data uncovered more nuanced user experiences that provide critical direction for future interface refinements. On the validation side, it's shown that 1) The demo effectively addresses small but time-consuming validation needs, 2) Transparency-focused design improves user experience quality, despite low feature utilization, and 3) Structured interface could potentially reduce bias perception and increase user confidence. On the discover side, it's suggested that 1) The name of "AI" has been confusing users, 2) The gap between LLM's promise of 'knowing everything' and users' need to understand its actual limits creates fundamental usability challenges, 3) Consider both distributed and global LLM workflows.

The culmination of these findings generated design guidelines structured along four critical dimensions: content , interaction, user journey, and positioning within the larger ecosystem.

Limitations & Future Directions

Despite the promising capabilities validated through this study, several limitations must be acknowledged. First, the limited number of participants (seven) means the findings may not be representative of the broader user population. While diverse participants were carefully selected, a larger sample size would provide more statistically significant results and potentially reveal additional patterns.

Second, the technical implementation did not fully consider more advanced LLM capabilities that have emerged recently, such as RAG and more advanced conversation management mechanisms. As the state of the art in LLM technology advances rapidly, the prototypes may not fully align with current LLM capabilities. This temporal limitation is inherent to research in fast-moving technological domains.

Third, many points discussed in the findings remain at a surface level. While various facets of the user experience were observed, due to scope constraints, it cannot be conclusively determined how specific interface elements influenced overall perceptions. The interplay between design choices and user expectations requires more granular investigation.

Looking forward, three key directions for future research emerge:

1. Expanding the participant range to include more diverse background would strengthen the validity and generalisability.
2. Better identification of current LLM capabilities and how they can be appropriately applied to shopping contexts would ensure that future designs leverage the most recent advancements in LLM related technologies.
3. Further identification and systematic study of the individual facets of LLM-mediated shopping experiences would provide more actionable design guidance. Isolating variables such as response structure, different interface design patterns would allow for more precise understanding of their impacts.

Reflection:

This research focused on interface design for LLMs; however, it hasn't addressed scenarios where LLMs might not be the appropriate solution, or how to train LLMs in human-centric ways. Up to this point, this thesis has not discussed how to improve the actual technology behind the interface. Some of the design suggestions are essentially solutions addressing the lack of transparency in the language model's training data, processing methods, and output stability. By the end of my project, my interest has extended into the area of model-tuning, with curiosity about how we could innovate by rethinking the development process of LLMs.

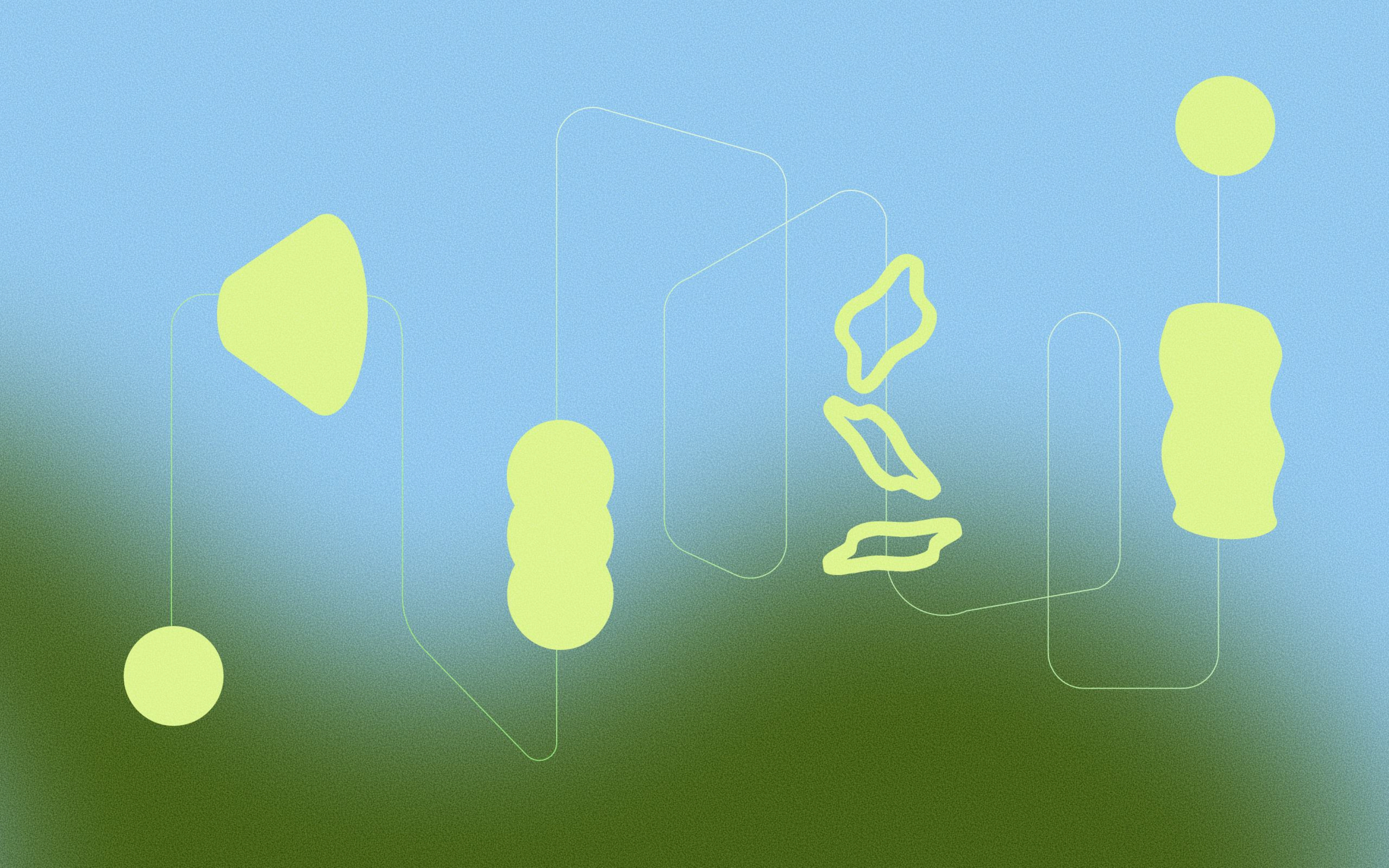
Another critical perspective is the balance between automation and autonomy. The adoption of LLM technologies often means automating parts of tasks that humans used to conduct - for instance, in this project, automating the process of looking through and synthesizing reviews.

However, humans sometimes find value and enjoyment in the process of tackling problems themselves, which raises the question of whether there should be boundaries for task automation.

These philosophical questions are part of my original motivation for continuing to explore the design of LLM-based systems and understanding their user experience. Looking back at my process, I think as a designer/engineer/researcher, the powerful tools I have for both creating and testing allow me not to rush into picking a side, but rather to take my time to create and test, and say 'this is what I see.' I can then build my suggestions upon these facts.

What is still lacking in my process, though, is more involvement of stakeholders in the beginning, where co-creation could have played a role in gathering different perspectives and ideas. This would allow me to more confidently transfer from the ideation stage to the validation stage, with an understanding of the organisational contexts.

Beyond that, through this project, I have developed an understanding of the technology that allows me to think of LLMs as building blocks of digital product design, and generate design suggestions that are grounded in user feedback and human-centric design principles. It has been a step I made toward better understanding of designing with LLM technologies - diving under the marketing hype, understanding the user experience reality, and defining what comes next from user perspective.



List of References

[A]

Adomavicius, G., Bauman, K., Tuzhilin, A., & Unger, M. (2022). Context-aware recommender systems: From foundations to recent developments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender systems handbook (pp. [page range]). Springer. https://doi.org/10.1007/978-1-0716-2197-4_6

Algolia. (2024). Decathlon Singapore: Driving 50% higher conversion rate with omnichannel, personalized search. https://www.algolia.com/customers/decathlon-singapore

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Anthropic. (n.d.). Model context protocol: Introduction. https://modelcontextprotocol.io/introduction

Anthropic. (2025). What are artifacts and how do I use them? Anthropic Help Center. https://support.anthropic.com/en/articles/9487310-what-are-artifacts-and-how-do-i-use-them

Arslan, M., & Cruz, C. (2024). Business-RAG: Information extraction for business insights. In 21st International Conference on Smart Business Technologies (pp. 88-94). SCITEPRESS-Science and Technology Publications.

[B]

BearingPoint. (2021, April). Retailer brand policies: Decathlon case study (1/2). https://www.bearingpoint.com/fr-fr/publications-evenements/blogs/blog-cpg-retail-luxury/retailer-brand-policies-decathlon-case-study-12/

Beverland, M. (2018, January). Brand management: Co-creating meaningful brands. https://openlibrary.org/books/OL27831405M/Brand_Management

Black, J. (2023, November). New ways generative AI can help you find holiday gifts. Google Blog. https://blog.google/products/shopping/google-search-generative-ai-holiday-gifts/

Bourgeois, O. (2024, September). An online shopping demo with Gemini and RAG. Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/an-online-shopping-demo-with-gemini-and-rag

Bradley, S., Ashley, S., & Park, Y. (2025, January). Build consumer confidence to boost your bottom line. Think with Google. https://www.thinkwithgoogle.com/consumer-insights/consumer-journey/build-confident-consumers-to-improve-metrics/

Braun, V., & Clarke, V. (2022, January). Thematic analysis: A practical guide. QMiP Bulletin, 1(33), 46-50. https://doi.org/10.53841/bpsqmip.2022.1.33.46

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. arXiv. https://arxiv.org/abs/2005.14165

[C]

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). Portrayals and perceptions of AI and why they matter. https://doi.org/10.17863/CAM.34502

Cheng, Z., Zhang, W., Chou, C.-C., Jau, Y.-Y., Pathak, A., Gao, P., & Batur, U. (2024, November). E-commerce product categorization with LLM-based dual-expert classification paradigm. In S. Kumar, V. Balachandran, C. Y. Park, W. Shi, S. A. Hayati, Y. Tsvetkov, N. Smith, H. Hajishirzi, D. Kang, & D. Jurgens (Eds.), Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (pp. 294-304). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.customnlp4u-1.22

Cho, A., Kim, G. C., Karpekov, A., Helbling, A., Wang, Z. J., Lee, S., Hoover, B., & Chau, D. H. (2024). Transformer explainer: Interactive learning of text-generative models. IEEE VIS Poster.

Cowan, N. (2010, February). The magical mystery four. Current Directions in Psychological Science, 19(1), 51-57. https://doi.org/10.1177/0963721409359277Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications.

Creusen, M. (2015). Consumer response to product form. In M. Luchs, K. Swan, & A. Griffin (Eds.), Design thinking: New product development essentials from the PDMA (pp. 303-317). John Wiley and Sons.

[D-E]

De Jong, C. W., Klemp, K., Mattie, E., & Goodwin, D. (2017, January). Ten principles for good design: Dieter Rams : the Jorrit Maan Collection. https://lib.ugent.be/en/catalog/rug01:002378057

Decathlon. (2022). Decathlon's story - our key dates. https://sustainability.decathlon.com/decathlons-story-our-key-dates

Decathlon. (2024). Decathlon new brand identity and positioning. https://www.decathlon-united.media/decathlon-new-identity

Decathlon. (2025). About Decathlon. https://www.decathlon.com/pages/about-decathlon?srsltid=AfmBOoqUSamWuhq7DkR2tos1boF5i7y2qH1dUNTm8xKikK8CF84vLjAG

Dharmadhikari, S. (2025, February). Pepsi's crystal clear failure: The story of crystal Pepsi. Cognitive Market Research. https://www.cognitivemarketresearch.com/blog/pepsis-crystal-clear-failure-the-story-of-crystal-pepsi?srsltid=AfmBOopUAPDLbnIHEdXD2oBX480TCOj_JhpJkelTQQqewkmUEiUIv2Ul

Epoch. (2025). Computation used to train notable artificial intelligence systems, by domain [Dataset]. Our World in Data. Retrieved May 19, 2025, from https://ourworldindata.org/grapher/artificial-intelligence-training-computation

Evans, M., Foxall, G., & Jamal, A. (2009). Consumer behaviour. Wiley.

[F-G]

Forristal, L. (2024, June). True Fit leverages generative AI to help online shoppers find clothes that fit. TechCrunch. https://techcrunch.com/2024/06/04/true-fit-generative-ai-feature-fit-hub/

Forristal, L. (2024, June). True Fit leverages generative AI to help online shoppers find clothes that fit. TechCrunch. https://techcrunch.com/2024/06/04/true-fit-generative-ai-feature-fit-hub/

Gonen, E. (2019, January). Tim Brown, Change by Design: How design thinking transforms organizations and inspires innovation (2009). Markets Globalization Development Review, 4(2). https://doi.org/10.23860/mgdr-2019-04-02-08

[H]

Hancock, P. M., Hilverman, C., Cook, S. W., & Halvorson, K. M. (2023, November). Emoji as gesture in digital communication: Emoji improve comprehension of indirect speech. Psychonomic Bulletin & Review, 31(3), 1335-1347. https://doi.org/10.3758/s13423-023-02411-1

Hekkert, P. P. M., & Van Dijk, M. B. (2011). Vision in design: A guidebook for innovators. BIS Publishers.

Heutz, M., & Rizi-Shorvon, E. (2024, November). Omnichannel best practices: What do winning retailers get right? Think with Google. https://www.thinkwithgoogle.com/intl/en-emea/consumer-insights/consumer-trends/best-omnichannel-brands-northern-europe/

Hmmh.DE. (2014). Decathlon search navigation. https://www.hmmh.de/en/projects/decathlon-search-navigation

Hohenstein, J., & Jung, M. F. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. Computers in Human Behavior, 106, 106190. https://api.semanticscholar.org/CorpusID:210501915

Hoyer, W., MacInnis, D., & Pieters, R. (2016). Consumer behavior. Cengage Learning.

[I]

IBM. (2024, January). Revolutionize retail with AI everywhere. https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-retail?utm_medium=Email&utm_source=External&utm_content=IBVWW&utm_term=20MFX&utm_id=1200-revolutionize-retail-AI-2024-press

Iyengar, S. S., & Lepper, M. R. (2000, December). When choice is demotivating: Can one desire too much of a good thing? Journal of Personality and Social Psychology, 79(6), 995-1006. https://doi.org/10.1037/0022-3514.79.6.995

[J-K]

Jamal, A., & Goode, M. M. (2001, December). Consumers and brands: A study of the impact of self-image congruence on brand preference and satisfaction. Marketing Intelligence & Planning, 19(7), 482-492. https://doi.org/10.1108/02634500110408286

Kim, J., Kim, J. H., Kim, C., & Park, J. (2023). Decisions with ChatGPT: Reexamining choice overload in ChatGPT recommendations. Journal of Retailing and Consumer Services, 75, Article 103494. https://doi.org/10.1016/j.jretconser.2023.103494

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300641

Krug, S., & Black, R. (2009, August). Don't make me think: A common sense approach to web usability. http://www.personal.kent.edu/~jinderhe/pdf%20handouts/don%27tmakemethink.pdf

[L]

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv. https://arxiv.org/abs/1910.13461

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://arxiv.org/abs/2005.11401

Loken, B., Barsalou, L. W., & Joiner, C. (2008). Categorization theory and research in consumer psychology: Category representation and category-based inference. https://api.semanticscholar.org/CorpusID:5914999

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5286-5297). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858288

[M-N]

Malhotra, N. K. (1982). Information load and consumer decision making. Journal of Consumer Research, 8(4), 419-430. https://doi.org/10.1086/208882

Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. Computers in Human Behavior, 128, Article 107093. https://doi.org/10.1016/j.chb.2021.107093

Norman, D. A. (1988, January). The design of everyday things. http://documents.irevues.inist.fr/bitstream/2042/30193/1/XX_CNE-Prospective_000738.pdf

[O-P]

OpenAI. (2024). Introducing canvas. https://openai.com/index/introducing-canvas/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). GPT-4 technical report. arXiv. https://arxiv.org/abs/2303.08774

Padurariu, O. (2025, January 8). Rufus – The blueprint. Seller Sessions. https://sellersessions.com/rufus-the-blueprint/

Park, Y. A., & Gretzel, U. (2010, January). Influence of consumers' online decision-making style on comparison shopping proneness and perceived usefulness of comparison shopping tools. Journal of Electronic Commerce Research, 11(4), 342-354. https://espace.library.uq.edu.au/view/UQ:326878/UQ326878_OA.pdf

Pillay, T. (2025, January 8). How OpenAI's Sam Altman is thinking about AGI and superintelligence in 2025. Time. https://time.com/7205596/sam-altman-superintelligence-agi/

Porter, M. E. (1985, January). Competitive advantage: Creating and sustaining superior performance. http://ci.nii.ac.jp/ncid/BA00852365

[R-S]

Rahwan, I. (2017, August). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(1), 5-14. https://doi.org/10.1007/s10676-017-9430-8

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. Natural Language Processing Journal, 6, 100056. https://doi.org/https://doi.org/10.1016/j.nlp.2024.100056

Solomon, M., Dahl, D., White, K., Zaichkowsky, J., & Polegato, R. (2013). Consumer behaviour: Buying, having, being. Pearson Learning Solutions.

Springer, A., & Whittaker, S. (2019). Progressive disclosure: Empirically motivated approaches to designing effective transparency. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 107-120). Association for Computing Machinery. https://doi.org/10.1145/3301275.3302322

Suomala, J. (2020). The consumer contextual decision-making model. Frontiers in Psychology, 11, Article 570430. https://doi.org/10.3389/fpsyg.2020.570430

[T]

Target. (2024, December). New GenAI-powered Bullseye gift finder makes it easy to pick out holiday gifts for kids of all ages. https://corporate.target.com/news-features/article/2024/12/bullseye-gift-finder

ThinkwithGoogle. (2024, August). Shopper types 01. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/shopping-research-before-purchase-statistics/

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In Conference on Spatial Information Theory. https://api.semanticscholar.org/CorpusID:2444850

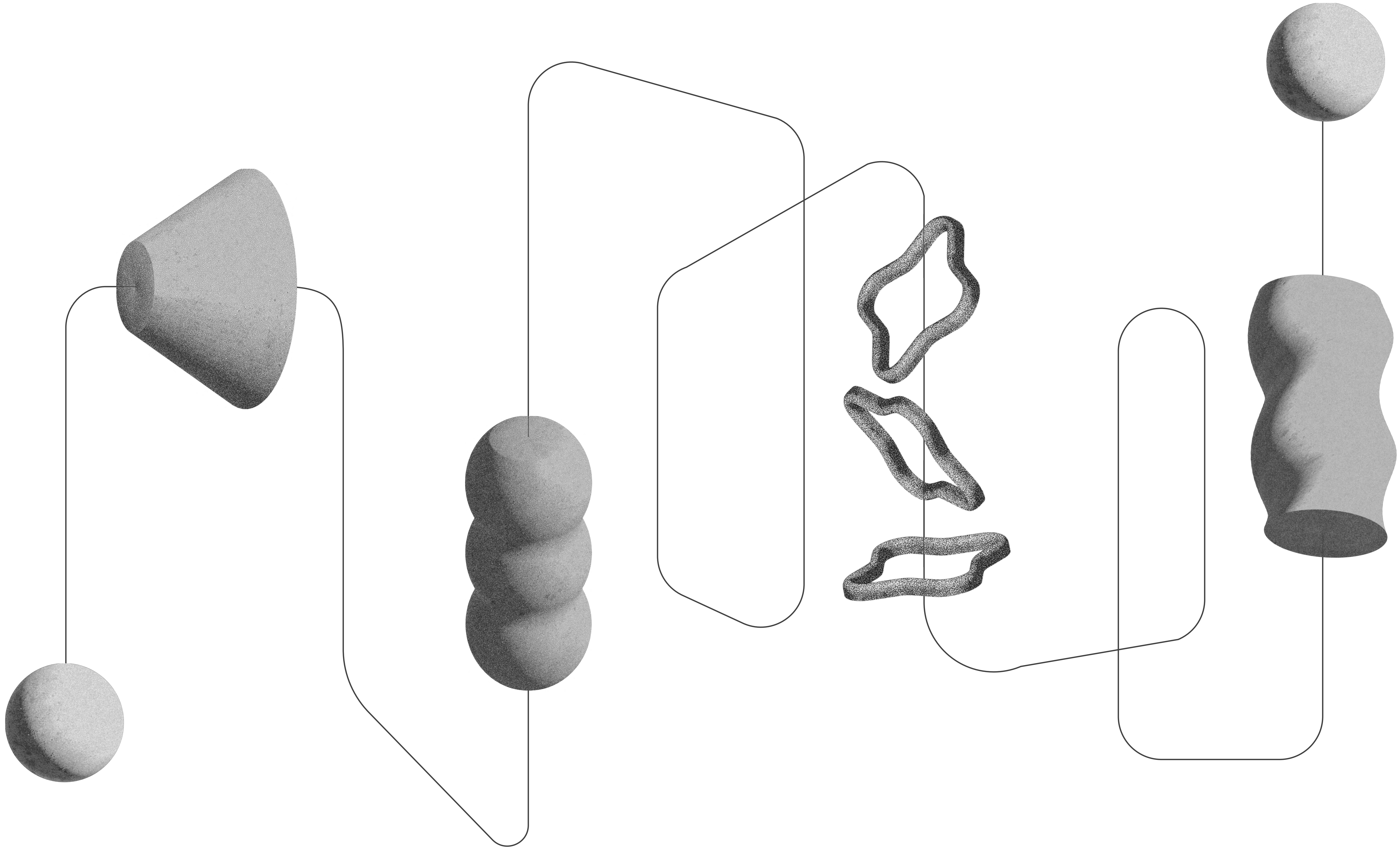
[V-W]

Van Boeijen, A., Daalhuizen, J., & Zijlstra, J. (2020, July). Delft design guide: Perspectives - models - approaches - methods. https://research.tudelft.nl/en/publications/delft-design-guide-perspectives-models-approaches-methods

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. arXiv. https://arxiv.org/abs/1706.03762

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2014, July). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. Motivation Science, 1(S), 19-42. https://doi.org/10.1037/2333-8113.1.s.19

Weil, E. (2023, March). ChatGPT is nothing like a human, says linguist Emily Bender. New York Magazine. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html



Bridging the Promise-Reality Gap

TU Delft

DESIGN FOR our future

IDE Master Graduation Project

Project team, procedural checks and Personal Project Brief

In this document the agreements made between student and supervisory team about the student's IDE Master Graduation Project are set out. This document may also include involvement of an external client, however does not cover any legal matters student and client (might) agree upon. Next to that, this document facilitates the required procedural checks:

- Student defines the team, what the student is going to do/deliver and how that will come about
- Chair of the supervisory team signs, to formally approve the project's setup / Project brief
- SSC E&SA (Shared Service Centre, Education & Student Affairs) report on the student's registration and study progress
- IDE's Board of Examiners confirms the proposed supervisory team on their eligibility, and whether the student is allowed to start the Graduation Project

STUDENT DATA & MASTER PROGRAMME

Complete all fields and indicate which master(s) you are in

Family name: Zhu

Initials: Q.Z.

Given name: Qiulin

Student number: 5578353

IDE master(s): IPD ☐ Dfl ☐ SPD ✔

2nd non-IDE master:

Individual programme *(date of approval)*:

Medisign ☐

HPM ☐

SUPERVISORY TEAM

Fill in he required information of supervisory team members. If applicable, company mentor is added as 2nd mentor

Chair	Dr. D.S. (Dave) Murray-Rust	dept./section	Human-Centered Design (HCD)
mentor	Dr. Ujwal Gadiraju	dept./section	EEMCS/EWI
2nd mentor	Abdelrahman (Abdo) Hassan		
client:	Decathlon Digital		
city:	Amsterdam	country:	the Netherlands
optional comments			

! Ensure a heterogeneous team. In case you wish to include team members from the same section, explain why.

! Chair should request the IDE Board of Examiners for approval when a non-IDE mentor is proposed. Include CV and motivation letter.

! 2nd mentor only applies when a client is involved.

APPROVAL OF CHAIR on PROJECT PROPOSAL / PROJECT BRIEF -> to be filled in **by the Chair** of the supervisory team

Sign for approval (Chair)

Dave Murray-Rust — Digitally signed by Dave Murray-Rust Date: 2024.11.20 18:14:12 +01'00'

Name ______ Date ______ Signature ______

CHECK ON STUDY PROGRESS

To be filled in **by SSC E&SA** (Shared Service Centre, Education & Student Affairs), after approval of the project brief by the chair.
The study progress will be checked for a 2nd time just before the green light meeting.

Master electives no. of EC accumulated in total ______ EC

Of which, taking conditional requirements into account, can be part of the exam programme ______ EC

☐	YES	all 1st year master courses passed
☐	NO	missing 1st year courses

Comments:

Sign for approval (SSC E&SA)

Name ______ Date ______ Signature ______

APPROVAL OF BOARD OF EXAMINERS IDE on SUPERVISORY TEAM -> to be checked and filled in by IDE's Board of Examiners

Does the composition of the Supervisory Team comply with regulations?

YES	☐	Supervisory Team approved
NO	☐	Supervisory Team not approved

Comments:

Based on study progress, students is ...

☐	**ALLOWED** to start the graduation project
☐	**NOT** allowed to start the graduation project

Comments:

Sign for approval (BoEx)

Name ______ Date ______ Signature ______

TU Delft

Personal Project Brief – IDE Master Graduation Project

Name student Qiulin Zhu **Student number** 5578353

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT

Complete all fields, keep information clear, specific and concise

Project title Exploring Meaningful Human Feedback for AI Recommendation System

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

Large Language Models (LLMs) have proven to be highly effective in predicting and generating text, making them valuable tools for a variety of tasks, including writing assistance and, more recently, customer service and product recommendation systems. These applications underscore the capability of LLMs to understand user intentions and facilitate more accurate product searches. To enhance these abilities, LLMs are often fine-tuned using specialized datasets that incorporate human feedback, ensuring the models adapt better to specific contexts.

However, the design and implementation of AI systems such as these extend beyond purely technical challenges, entering the realm of sociotechnical systems. This shift introduces complexities that encompass both technological and social dimensions. A prime example is Reinforcement Learning from Human Feedback (RLHF), a model tuning technique where ongoing discussions focus on critical concerns, such as the selection of representative human annotators, the difficulty in ensuring the quality of feedback data, and the underlying values the model should reflect (Figure 1).

In collaboration with Decathlon Digital, my project seeks to engage a broader range of stakeholders involved in the development of Decathlon's AI recommendation system, including consumers, sports experts, and product developers. This participatory approach aims to integrate community knowledge and values more deeply into the AI system, fostering the creation of a recommendation model that better reflects the diverse perspectives and needs of its users.

→ *space available for images / figures on next page*

introduction (continued): space for images

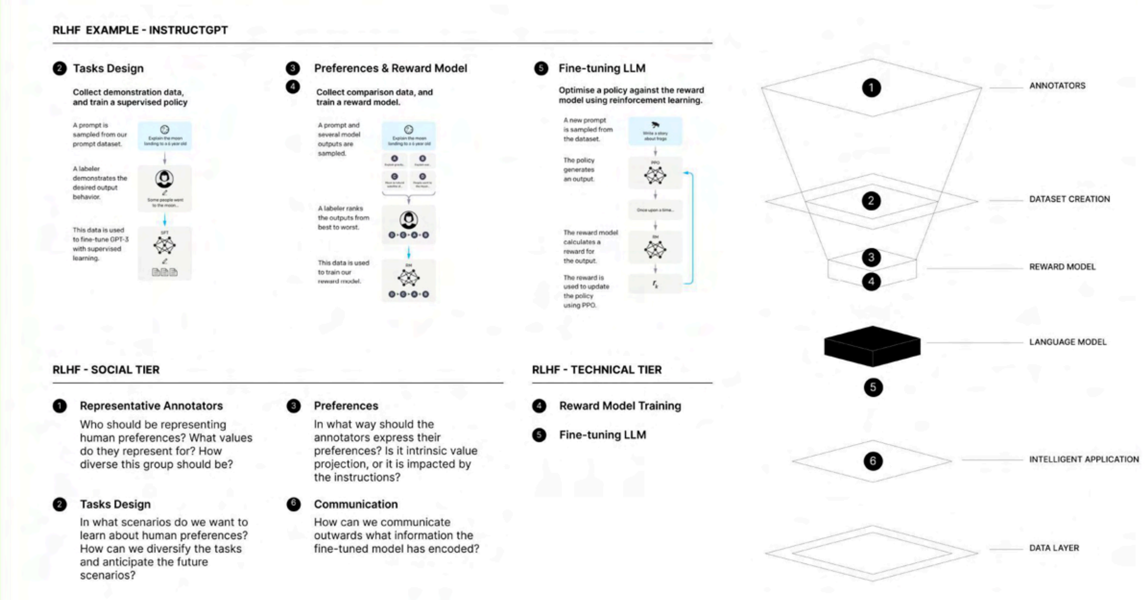


image / figure 1

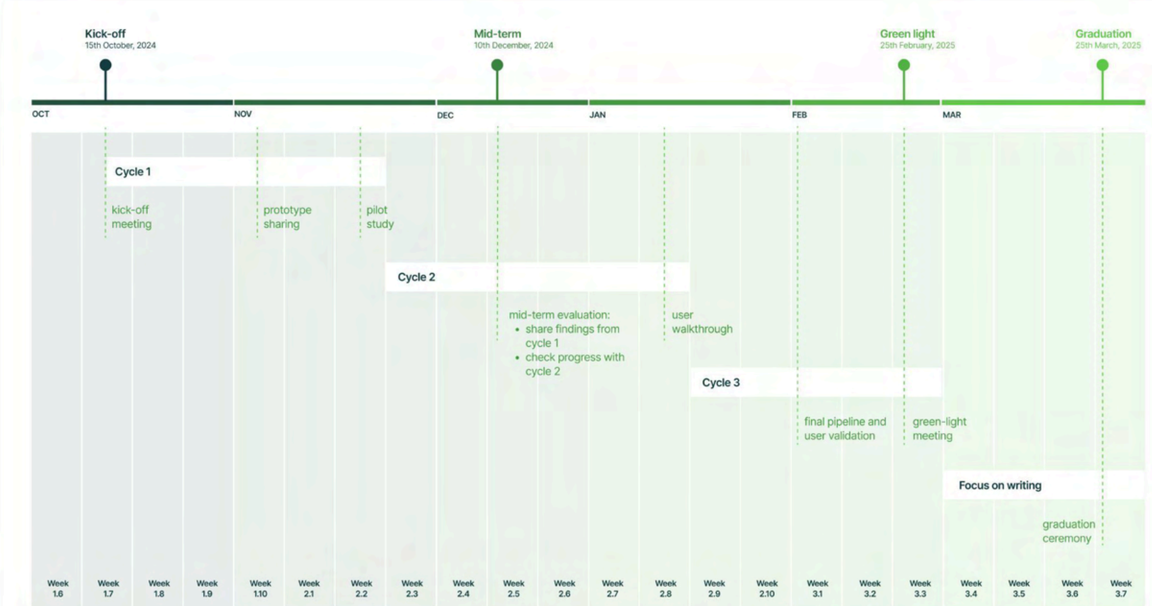


image / figure 2

DESIGN FOR our future

TUDelft

Personal Project Brief – IDE Master Graduation Project

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

The vision of incorporating community-driven human feedback into AI systems is undeniably ambitious, requiring efforts that extend beyond the immediate scope of this project. The core issue this project aims to address is the challenge of defining what constitutes "meaningful" feedback within the context of Decathlon's AI product recommendation system. While traditional feedback mechanisms often focus on simple preference-based responses, this project seeks to explore more sophisticated feedback formats that offer deeper, more nuanced insights into user experiences.

Some key aspects of this challenge include:

1. Exploring how to gather feedback that not only captures general preferences on outputs but also enables the LLM to interpret needs at a more complex, context-specific level.
2. Developing interaction mediums that facilitate meaningful feedback sharing, fostering constructive input to inspire positive AI system development.
3. Investigating how this feedback can be effectively integrated into the AI system's development process to ensure continuous improvement.

Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Create an interactive artifact that enables meaningful feedback sharing, leveraging community-driven and contextual knowledge to inform AI development.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

I will approach the project using a creative, hands-on design exploration process, with a strong emphasis on iterative prototyping and user studies. The exploration will follow three design cycles, each cycle progressively deepening both the understanding of the problem and the refinement of the solution (Figure 2).

The design solution will be developed through digital artifact prototyping, incorporating human interaction with LLMs. This will involve exploring various elements—such as data points, interaction patterns, and visual storytelling—to facilitate meaningful engagement and interaction.

Additionally, I will engage with the multistakeholder community through interviews and workshops to understand their expectations and experiences. This will allow for reflection on their interactions with LLMs and the overall development lifecycle.

Project planning and key moments

*To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a **kick-off meeting**, **mid-term evaluation meeting**, **green light meeting** and **graduation ceremony**. Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting	15th October, 2024
Mid-term evaluation	10th December, 2024
Green light meeting	25th February, 2025
Graduation ceremony	25th March, 2025

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	✔
For how many project weeks	20
Number of project days per week	4

Comments:
Allocate one day per week for electives and career development.

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

Reflecting on the courses I've taken over the past three semesters, I've developed a deep interest in exploring AI from various perspectives. Early on, in the Roadmapping course with IBM, I began cultivating my interest in responsible AI development. Later, in the AI&Society course, I expanded my perspective beyond user-centric, learning to evaluate technology through considering multiple stakeholder groups. In the Explore Design Intelligence course, I gained skills in analyzing and visualizing natural language datasets, and in Crowd Computing System Design, I learned the complexities of designing tasks for collecting human feedback data.

During my internship with Decathlon's AI Innovation and Trust team, I gained valuable insight into how AI can be applied in real-world contexts. While technical aspects remain central to innovation, I realized that imagination plays a crucial role not only in algorithm design but also in addressing the complexities and interconnectedness of human behavior in the real world. I believe human feedback perspective offers a starting point for me in the full AI development cycle, where I could further develop my skills in helping organizations navigate AI development with a designer's empathy, human-centered design methodologies, and creative problem-solving.