UPPSALA
UNIVERSITET

UPTEC STS 25016
Examensarbete 30 hp
Juni 2025

Beyond the Hype: Augmenting Knowledge Work with Agentic AI

A Service Design Approach to Enterprise Transformation

Alma Lundberg

Civilingenjörsprogrammet i System i Teknik och Samhälle

UPPSALA
UNIVERSITET

Beyond the Hype: Augmenting Knowledge Work with Agentic AI

Alma Lundberg

Abstract

The rapid emergence of generative AI marks a paradigm shift for knowledge-intensive organizations, but many enterprises still face challenges in translating this potential into tangible value. This thesis explores how generative AI can strengthen knowledge-intensive workflows, drawing on a case study within the product management function of a large enterprise. Using a service design perspective, the research employs qualitative, user-centered methods to identify recurring challenges and evaluate high-impact opportunities for AI intervention.

Findings indicate that while generative AI can streamline tasks such as information synthesis, summarization, and competitor research, lasting value requires more than prompt-based automation. Agentic systems, capable of orchestrating planning, reasoning, and tool use, offer a more robust pathway for integrating AI into real workflows. The development and validation of a agentic research assistant prototype illustrated both the potential and the organizational challenges of implementing agentic AI.

The study concludes that participatory, iterative approaches are critical if generative AI is to successfully augment working practices in knowledge-intensive organizations. By integrating AI into existing workflows and aligning with user needs, organizations can realize its full potential and lay the groundwork for successful AI transformation. This research provides actionable guidance for practitioners navigating the rapidly evolving AI landscape and demonstrates how agentic AI can deliver real value in enterprise contexts.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Tom Widegren Ämnesgranskare: Jessica Lindblom

Examinator: Elísabet Andrésdóttir

Populärvetenskaplig sammanfattning

Artificiell intelligens (AI) diskuteras idag som en teknik med potential att förändra arbetslivet i grunden, inte minst för dem som arbetar med information, analys och beslutsfattande, något som brukar kallas kunskapsarbete. Generativ AI, som kan skapa text, data och sammanfatta information, framställs ofta som nästa stora genombrott. Men samtidigt är det långt ifrån självklart hur denna teknik faktiskt kan ge nytta på jobbet, och många satsningar tenderar att fastna i småskaliga försök utan större genomslag.

I praktiken visar det sig ofta att arbetsvardagen är mer komplex än tekniken först ger sken av. Många upplever att de digitala verktygen visserligen kan svara på frågor, men sällan är anpassade till verkliga arbetsflöden, eller de samarbetsformer och rutiner som finns i en organisation. Risken är att tekniken snarare bidrar till ännu mer fragmenterad information och nya sorters frustration.

Det här arbetet har undersökt hur AI faktiskt kan stärka och utveckla arbetsprocesser i en större organisation, med fokus på kunskapsintensiva roller där analys, informationssökning och samarbete är centrala delar av jobbet. Utgångspunkten har varit att förstå de utmaningar och behov som medarbetare själva lyfter, snarare än att börja med tekniska möjligheter. För att göra detta har jag använt ett arbetssätt som kallas tjänstedesign, en metod där man tillsammans med användarna utforskar hur arbetsuppgifter faktiskt utförs, identifierar problem och tillsammans skapar nya lösningar som passar in i den dagliga verksamheten.

En särskild del av studien handlar om den nya generationen AI-system, så kallade agentiska AI-lösningar. Till skillnad från traditionella AI-verktyg, som ofta svarar på enskilda frågor eller hjälper till med avgränsade uppgifter, kan agentiska system själva planera, söka, sammanfatta och bearbeta information över flera steg och därmed bli en mer aktiv samarbetspartner för människor. Tanken är inte att ersätta människor, utan att frigöra tid och energi från monotona och tidskrävande arbetsmoment, så att människor kan fokusera på mer värdeskapande och utvecklande uppgifter.

Under arbetets gång har jag identifierat och analyserat de arbetsuppgifter som upplevs som mest tidskrävande, splittrade eller manuella, till exempel att samla in och analysera marknads- och konkurrentinformation. Genom att intervjua användare och kartlägga deras arbetsflöden, har jag tillsammans med dem utvecklat och testat en prototyp av en AI-baserad assistent. Denna assistent kan inte bara hämta information snabbare, utan

även hjälpa till att organisera, sammanfatta och presentera insikter, alltid med användaren i förarsätet.

Resultaten visar att AI har stor potential att effektivisera arbetsuppgifter som idag upplevs som svåröverblickbara och onödigt manuella. Men avgörande är att nyttan uppstår först när AI-lösningar utvecklas tillsammans med dem som ska använda dem, och anpassas till verkliga behov och arbetsmönster. Det är också viktigt att behålla mänsklig kontroll och ansvar i arbetsprocesserna, så att AI blir ett stöd och ett verktyg i arbetet, snarare än något som agerar självständigt utan insyn.

Sammanfattningsvis visar studien att framtidens arbetsliv kan bli både mer effektivt och stimulerande med hjälp av AI, men bara om tekniken införs på ett sätt som utgår från verkliga arbetsprocesser, användarnas perspektiv och ett nära samspel mellan människa och maskin. Genom att kombinera tjänstedesign med den senaste AI-tekniken skapas nya möjligheter för arbetsplatser att utvecklas, samtidigt som mänsklig erfarenhet och omdöme står kvar i centrum.

Acknowledgements

This thesis would not have been possible without the support and encouragement of many people. First and foremost, I would like to thank all participants in the study who generously shared their experiences, insights, and time. Your openness and willingness to discuss both challenges and opportunities have been crucial for the quality and relevance of this work.

I would especially like to thank TW, who initiated this project and has been an inspiring mentor throughout my journey. Your forward-thinking mindset and encouragement have played a key role not only in shaping this thesis but also in advancing the broader conversation around AI within the organization. I am truly grateful for your trust, vision, and continued support.

My deepest thanks also go to my academic supervisor, Jessica Lindblom, for your outstanding guidance and expertise. Your encouragement, deep theoretical knowledge, and thoughtful feedback have been invaluable. I have greatly appreciated our discussions and your unwavering belief in my ability to complete this work.

A heartfelt thank you to AR for your invaluable AI expertise, guidance, and collaboration. Your openness and support have made a real difference and enabled me to connect with colleagues across the organization.

Finally, I would also like to thank Amir Elion for being a sharp and inspiring sounding board during the spring, and for facilitating important dialogues and engagement within the project.

Without your contributions, this thesis would not have been possible.

- Alma

Contents

1 Introduction

The world is witnessing a technological transformation at a pace and scale unseen since the advent of the internet. Over the past year, the rise of generative AI has moved with breathtaking speed, fundamentally disrupting not just the technology sector but the very fabric of how organizations create, process, and act on knowledge [1, 2, 3]. No longer confined to research labs or speculative headlines, generative AI is being woven directly into the workflows of enterprises, reshaping decision-making, collaboration, and innovation across every major industry [4, 5, 3]. This surge is not a matter of incremental improvement; it represents a true paradigm shift. Analysts and practitioners alike describe this as an "AI moment," where the collective ambitions, investments, and expectations surrounding artificial intelligence are converging into a generational turning point for business and society [3, 6, 7, 8].

Yet, as organizations race to capitalize on generative AI, the reality inside enterprises is more complex than the hype suggests. Despite rapid adoption and enormous investment, many implementations stall at the proof-of-concept stage or struggle to deliver measurable value. Empirical studies and large-scale surveys consistently highlight a widening gap between the transformative promise of AI and its realized impact on productivity, coordination, and meaningful work [9, 10, 11, 12, 13, 14]. Instead of seamless augmentation, many knowledge workers experience fragmented workflows, new forms of cognitive overload, and challenges integrating AI into established routines. This has shifted the discourse away from AI as a replacement for human labor, toward a more realistic vision of AI as an augmenting collaborator: one that enhances, rather than diminishes human judgment, creativity, and expertise [15, 16].

It is in this shifting landscape that agentic systems have re-emerged as a vital conceptual and technical development. While the roots of agentic architectures stretch back decades in AI research [17, 18, 19, 20, 21], it is the convergence with generative AI, specifically large language models and their orchestration in tool-using, planning-capable systems, that is unlocking new possibilities [22, 23]. Agentic systems represent the most ambitious attempt yet to move beyond static, prompt-based automation, instead orchestrating AI components that can autonomously plan, reason, and interact with complex digital environments. In this context, agentic systems are not a separate field but a critical extension of the generative AI revolution, offering a path to sustained, context-aware augmentation of knowledge work.

The impact of these advances is being felt most acutely in knowledge-intensive roles, those where value is created by synthesizing, interpreting, and acting on information rather than by routine or manual labor. Knowledge work is central to productivity in modern enterprises, yet remains resistant to simple digitization [24, 25, 13, 14]. Decades of digital transformation have delivered new tools and platforms, but persistent barriers such as fragmented information, manual coordination, and cognitive overload remain unresolved [26, 9]. The so-called "productivity paradox" endures: organizations invest heavily in technology, yet struggle to realize proportional gains in knowledge worker effectiveness [9].

To break through this paradox, there is growing recognition that transformation must be rooted in a nuanced understanding of actual work practices, routines, and user needs [27, 28]. Service design provides such a framework, emphasizing contextual immersion, co-creation, and iterative prototyping [28]. In the context of generative and agentic AI, this approach is critical for ensuring that new tools are embedded meaningfully within the realities of organizational life reducing the risk of fragmentation and technology-driven failure, while fostering adoption, trust, and sustained value [15, 16].

This study addresses these challenges by empirically investigating the conditions and opportunities for embedding generative AI within the knowledge work of a large enterprise. Specifically, the research is situated within the product management function of a major business area, characterized by high cognitive demands and cross-functional collaboration. Through direct engagement with real workflows and ongoing stakeholder participation, the study applies service design principles to analyze pain points, evaluate AI opportunities, and prototype solutions that reflect both user needs and organizational context.

1.1 Aim and objectives

The aim of this study is to investigate how generative AI can augment knowledge work within a large enterprise context using a service design perspective. This aim is pursued through the following objectives:

1. *To identify recurring workflow challenges and user needs in knowledge work that may be addressable through generative AI.*
2. *To evaluate and prioritize opportunity areas for generative AI intervention in knowledge work.*
3. *To design and develop an advanced prototype that demonstrates how generative AI can be meaningfully embedded into a selected knowledge work process.*

1.2 Collaboration and Case Context

This study was conducted in close collaboration with a large global enterprise which was anonymized at the sponsor's request. It was selected for its active and ongoing exploration of generative AI in knowledge-intensive workflows. The research was situated specifically within one major business area's product management function, a unit characterized by high cognitive demands, frequent cross-functional collaboration, and ongoing responsibility for market- and customer-facing products and services.

This setting provided the opportunity to investigate how generative AI could meaningfully augment complex knowledge work in a real-world, enterprise-scale context. Collaboration with internal stakeholders enabled direct access to live workflows, internal artifacts, and organizational routines, as well as sustained engagement with the primary user group throughout the research process. By embedding the study within day-to-day activities and decision points, the research was able to capture situated pain points, surface user needs, and co-design and evaluate solutions in close alignment with real operational priorities. In line with the human-centered service design methodology, this collaborative case approach ensured that both problem discovery and solution prototyping remained grounded in the actual context of use, rather than abstract or hypothetical scenarios. The findings and conclusions of the study are thus specific to the participating business area and its product management function, but offer broader lessons for organizations pursuing enterprise AI adoption in knowledge-intensive workflows.

2 Background

2.1 Knowledge Work in Modern Organizations

Knowledge work has become the cornerstone of productivity and value creation in modern organizations, particularly in industries where problem-solving, analysis, and creative judgment are central to competitive advantage. The term "knowledge worker" refers to individuals whose primary resource is knowledge and who add value through the manipulation, generation, and dissemination of information [24]. In the literature, this type of work is also commonly referred to as white-collar work, in contrast to blue-collar work, which typically involves manual or routine physical labor. These workers generally have a high degree of autonomy, use both tacit and explicit expertise, and often need to adapt to changing situations, frequently working together with others [24, 29].

What makes knowledge work distinct is its cognitive intensity, its reliance on information as the main input, and its importance for post-industrial economies [25]. Unlike blue-collar work, knowledge work involves ambiguity, non-repetitive tasks, and ongoing learning. Professions such as product management, scientific research, and software engineering exemplify these dynamics [13]. In these roles, workers not only carry out tasks but also interpret, shape, and sometimes even redefine their own objectives, often in settings where information is incomplete or changes quickly [24, 25].

2.1.1 From Industrial Efficiency to Knowledge Work Complexity

Historically, the management of productivity has shifted from optimizing manual labor through standardized methods in the industrial era, toward addressing the more complex challenge of improving knowledge worker productivity [24]. While twentieth-century organizations focused on maximizing efficiency through standardization and the use of capital equipment, today's organizations see their core assets as the expertise, skills, networks of their workforce [24, 14]. However, even with the widespread adoption of digital tools, from enterprise platforms to online collaboration systems, there is still a clear gap between the promise of digital transformation and actual improvements in knowledge worker productivity [14, 16].

A central theoretical issue is the nature of knowledge itself. Early approaches conceptu-

alized knowledge as an asset to be codified, stored, and transferred within organizations. Later work, however, emphasizes the enacted and collective nature of "knowing in practice"; the idea that expertise and meaning are continually constructed and negotiated through everyday activities, particularly in distributed or agile organizational contexts [29]. This view highlights that knowledge work is deeply embedded in social practices, organizational routines, and informal networks, which makes managing and measuring it more complex than manual tasks [29, 14].

Despite its importance, knowledge work still faces persistent challenges. Empirical studies have consistently identified pain points such as fragmented information landscapes, manual data handling, cognitive overload from excessive digital inputs, and coordination barriers in cross-functional teams [14, 13, 29]. These problems are becoming more pronounced as organizations face faster product cycles, greater demands for agility, and more distributed or remote work [16, 30]. For example, recent studies show that knowledge workers often feel that their productivity is limited not by the amount of available information, but by the difficulty of making sense of it and acting on it in collaborative settings [13, 30].

Moreover, there is also an ongoing "productivity paradox": even as digital tools become more sophisticated, actual productivity gains in organizational productivity remain limited, especially in complex or innovation-driven environments [14, 30]. This gap is often explained by continued reliance on manual coordination, the difficulty of capturing and sharing tacit knowledge, and the rising demands of communication and sense-making in distributed teams [29, 16].

2.1.2 The Rise of Human-AI Collaboration

In light of these persistent challenges, the emergence of advanced artificial intelligence (AI) has brought renewed attention to the future of knowledge work. Unlike earlier waves of digital tools, contemporary AI has the potential to directly support or reshape core activities in knowledge workflows [31, 13, 30]. Rather than replacing professionals, current research and practice emphasize collaborative models, where AI operates as a "copilot" or workflow agent that complements human expertise [16, 30]. *Human–AI collaboration* in this context refers to the deliberate integration of AI systems into human workflows, enabling machines to handle structured, automatable tasks while leaving hu-

mans responsible for higher-order judgment, critical thinking, contextual understanding, and creativity [16, 31]. Studies suggest this model can improve productivity, especially for repetitive or low-value tasks, while preserving human agency in decision-making and problem-solving [16, 13].

Recent research shows that, in practice, AI is most often integrated as an effort-saving tool that automates routine and repetitive tasks, but always under human supervision [13]. While this approach offers efficiency gains, it simultaneously raises concerns about the potential for deskilling, reduced engagement, and over-reliance [13, 14]. Two primary patterns of human-AI collaboration are now emerging: AI as a copilot working interactively alongside humans, and as a workflow agent automating multi-step processes with the results reviewed and contextualized by human experts [30]. These approaches aims to ensure that humans remain embedded in key stages of the workflow, such as validation, interpretation, and creativity, so that AI augments rather than substitutes human capabilities [31].

However, the dominant framing of “human-in-the-loop” is increasingly being challenged as overly simplistic and technologically driven. Recent literature instead advocates for a “human-in-control” approach, emphasizing the need for systems that increase automation without diminishing human responsibility or oversight [15, 32]. This reframing acknowledges that effective AI integration requires more than just periodic human intervention. It demands sustained control, transparency, and the empowerment of human decision-makers throughout the process.

Addressing the associated challenges, such as building trust, maintaining transparency, and supporting ongoing skill development, will be essential as organizations adopt more advanced and embedded forms of AI-augmented work.

2.2 Generative AI

AI refers to systems capable of performing tasks that typically require human cognitive abilities, such as perception, reasoning, and decision-making. A key paradigm shift in AI occurred with the rise of machine learning (ML), a method that allows computers to learn patterns from data rather than following explicitly programmed rules. Arthur Samuel, an early pioneer in the field, famously defined ML as programming a digital computer "to behave in a way which, if done by human beings or animals, would be

described as involving the process of learning" [33, p. 210]. His self-learning checkers program was among the first demonstrations of how machines could improve through experience. Today, ML forms the backbone of most AI systems, enabling capabilities across fields such as computer vision, recommendation systems, and natural language processing [34].

Within this data-driven paradigm, generative AI has emerged as a class of models capable of producing new content, including text, images, and code, by learning the statistical structure of training data. Rather than focusing on recognition or classification, generative systems are designed to synthesize coherent and contextually relevant outputs. A central area of application is Natural Language Processing (NLP), which enables machines to process and generate human language. [35, 36, 37] One of the most significant breakthroughs in this area is the emergence of Large Language Models (LLMs), powerful generative systems that have redefined the landscape of language-based AI. The next section examines these models in greater detail.

2.2.1 Large Language Models: Definition and Emergent Abilities

Large Language Models (LLMs) are a class of machine learning systems trained on massive corpora of text data to predict and generate human-like language. They are trained on extremely large collections of text, such as books, websites, and articles, and use this data to learn and predict patterns in how language is used. These models are built with a very high number of parameters, often in the tens or even hundreds of billions, which are adjusted during training to help the model make better predictions. As the size of these models increases, they tend to develop more human-like skills. Recent surveys describe this trend as a shift from narrow, task-specific tools to more general-purpose systems that can handle a wide variety of language tasks across different domains [35, 36, 38].

The underlying architecture of LLMs is the Transformer, a neural network design that uses a mechanism called self-attention to evaluate how each word in a sentence relates to others [39]. Unlike earlier models that processed words sequentially, the Transformer architecture use a mechanism called self-attention to evaluate the importance of every word in a sentence relative to every other word. This allows them to capture context and meaning more effectively [35, 40]. As researchers began to increase the size of these

models and the amount of data and computation used during training, they observed that performance improved in a surprisingly predictable way. This pattern are now formalized through what are known as "scaling laws". These describe how model performance improves in relation to size, data, and compute, and have made it possible to anticipate gains from continued investment in larger systems [35, 36].

Emergent Abilities and In-Context Learning

The effects of scaling are not only quantitative. Once models reach a certain size, they begin to exhibit *emergent abilities*, which are abilities that do not appear in smaller models. One of the most well-documented examples is *in-context learning*. This means that the model can learn how to perform a task by simply being shown examples in the input, without any additional training or updates to its internal parameters. For instance, if the model is shown a few question and answer pairs, it can often recognize the pattern and answer a new question in the same format [35, 36, 38, 41].

This ability allows the model to handle new tasks with a small number of examples (few-shot) or none at all (zero-shot), as long as the task is described clearly in the input. In a few-shot setting, the input includes several examples that demonstrate how the task should be done. In a zero-shot setting, the model receives only an instruction or task description, without examples. Other emergent abilities include the ability to follow instructions in natural language and to solve problems that require several steps of reasoning [35].

These abilities were first systematically observed in the GPT-3 model [41] and marked a major shift in how researchers understood the potential of LLMs. They showed that models could perform entirely new tasks by recognizing patterns at inference time, which is the stage when the model generates outputs, rather than through additional training. Inference-time pattern recognition refers to this process, where the model relies solely on the information provided in the prompt to infer what it is being asked to do. It does this without accessing updated internal knowledge or undergoing fine-tuning. Notably, these abilities tend to emerge suddenly as model size increases, rather than developing gradually. This kind of abrupt emergence is considered a defining feature of large language models trained at very large scale. In these cases, the combination of model size, training data, and computational resources gives rise to capabilities that do not appear in smaller or less complex models [36, 38, 41].

Reasoning and Chain-of-Thought

The "reasoning" capability in LLMs stands out as one of the most surprising and consequential emergent behaviors. While LLMs are fundamentally trained as next-token predictors, empirical studies show that, once scaled sufficiently, they begin to exhibit behaviors that resemble logical inference, step-by-step problem solving, and even basic forms of planning. These capabilities are not explicitly programmed but emerge as a result of increasing model size, training on diverse datasets, and the expressive capacity of the Transformer architecture [35, 36, 38].

One of the most widely studied examples of this is chain-of-thought (CoT) reasoning. This refers to the model's ability to "think out loud" by generating a series of intermediate steps before giving a final answer. Instead of responding immediately with a solution, the model first explains its reasoning behavior. This simple change in how we prompt the model, known as CoT prompting, has been shown to dramatically improve performance on tasks that require logical reasoning, math, and problem-solving [42]. For example, when tested on a challenging dataset of math word problems, a very large model called PaLM was able to outperform earlier models by a wide margin simply because it was prompted to show its reasoning process step-by-step [42].

Importantly, this type of reasoning does not appear in smaller models. Researchers have found that chain-of-thought prompting only works reliably when the model is large enough, typically over 100 billion parameters [35, 42]. This pattern has also been observed in broader evaluations of large language models, where CoT reasoning is seen as one of several high-level capabilities that emerge only at scale [35, 38]

Since then, researchers have developed several extensions of CoT prompting that further improve model performance on complex reasoning tasks. These include self-consistency prompting, where the model generates multiple reasoning paths and selects the most common answer; scratchpad prompting, where intermediate steps are mixed with computations to help the model "think (calculating) while working"; and tree-of-thought prompting, which enables the model to explore, evaluate, and revise alternative solution paths [35, 38]. Each of these builds on the core insight that how a task is presented to the model, that is, how it is prompted, strongly influences whether reasoning behavior emerges at all. These techniques reflect a broader shift in focus from just training larger models to designing better ways of interacting with them, an area known as prompt engineering.

2.2.2 Prompting Strategies for Structured Reasoning

The previous section introduced CoT prompting as a mechanism through which LLMs can be guided to exhibit multi-step reasoning. However, CoT is only one example of a broader family of prompting strategies that have been developed to elicit more structured and reliable behavior from LLMs. Prompting has evolved from a formatting trick into a key interface for activating and modulating model capabilities across tasks. Recent surveys highlight prompting as a central technique in aligning LLMs to human goals, particularly in cases where direct training or fine-tuning is impractical [35, 36].

Prompting strategies differ based on how much contextual information they provide, the structure of the prompt, and the assumptions they make about model generalization. Table 1 outlines common prompting setups, ranging from simple zero-shot and few-shot formats to more interactive and iterative forms like multi-turn prompting. These vary in complexity and use case, from single-turn tasks to more advanced setups involving feedback loops and evolving context. Multi-turn prompting involves sequential prompts, often with tool feedback or evolving task state. An increasingly prominent example is Retrieval-Augmented Generation (RAG). In RAG, the language model interacts with an external retrieval system, such as a search engine or document database, to supplement its responses with up-to-date or domain-specific information. When a query is posed, relevant documents are retrieved and incorporated into the prompt for subsequent model responses, allowing the LLM to ground its output in external sources rather than relying solely on its internal knowledge. This hybrid approach not only improves factual accuracy and transparency, but also enables more flexible and updatable knowledge access, making it well suited for knowledge-intensive tasks [43, 44].

Table 1: Common prompting setups for Large Language Models.

Strategy	Description	Remarks
Zero-shot prompting	The model is given only an instruction without examples, relying entirely on prior knowledge.	Low setup cost; can underperform on reasoning-heavy tasks [36]
Few-shot prompting	Provides a few examples within the prompt to demonstrate input–output patterns.	Effective for in-context learning; sensitive to example phrasing/order [41]
Instruction prompting	Uses natural cues or identity framing to align model behavior with user intent.	Boosts general task performance; effective with instruction-tuned models [38]
Single-turn prompting	Full task is handled in a single prompt-response cycle.	Simple to implement; lacks iterative refinement capability [36]
Multi-turn prompting	Involves sequential prompts, often with tool feedback or evolving task state.	Enables interaction and planning; used in agents and RAG pipelines [35]

As discussed previously, CoT prompting is a strategy where the model is guided to articulate intermediate reasoning steps before arriving at a final answer. While CoT improves performance on tasks involving arithmetic, logic, and symbolic reasoning, it also exhibits limitations such as inconsistency, omissions, or shallow reasoning. In response, researchers have developed a range of enhanced prompting strategies, many of which explicitly aim to improve reasoning depth, structure, and accuracy. Table 2 presents an overview of enhanced prompting strategies that aim to improve reasoning performance. These include techniques like Self-Consistency and Plan-and-Solve, which add robustness through sampling or structured decomposition, as well as more experimental strategies like Tree-of-Thought (ToT) and Graph-of-Thought (GoT), which organize reasoning as a branching or recombination process.

These prompting strategies have been developed to address key limitations in standard CoT reasoning, such as hallucinated logic, incomplete intermediate steps, or the inability to revisit and revise prior reasoning paths. As illustrated in Figure 1, prompting strategies have progressed from simple linear instruction-based formats such as zero-shot and few-shot prompting toward more structured and dynamic methods like CoT, ToT, and GoT. The figure visualizes this evolution as a spectrum, moving from simple in-context learning toward increasingly exploratory and multi-path reasoning strategies, reflecting an increasing awareness of the complexity required to guide LLMs through robust reasoning processes.

Table 2: Prompting strategies for structured reasoning in LLMs.

Strategy	Description	Remarks
Chain-of-Thought (CoT)	Guides the model to produce intermediate reasoning steps before the final answer.	Improves logical accuracy; effective only at large scale [42, 35]
Self-Consistency	Samples multiple CoT reasoning paths and selects the most common answer.	Reduces output variance; increases compute cost [35]
Scratchpad prompting	Inserts reasoning "workspace" for the model to track intermediate variables or states.	Improves symbolic tasks; requires prompt formatting care [40]
Plan-and-Solve	Separates problem-solving into a planning phase and an execution phase.	Reduces missing-step errors; strong zero-shot performance [45]
Expert prompting	Frames the model as a domain expert, shaping output quality and tone.	Boosts domain fluency; performance varies by prompt detail [46]
Tree-of-Thought (ToT)	Structures reasoning as a tree of options that can be explored and evaluated.	Supports search-based problem solving; requires orchestration [47]
Graph-of-Thought (GoT)	Generalizes ToT by allowing reasoning paths to merge, refine, or recombine.	Enables flexible multi-path reasoning; promising but still experimental [48]

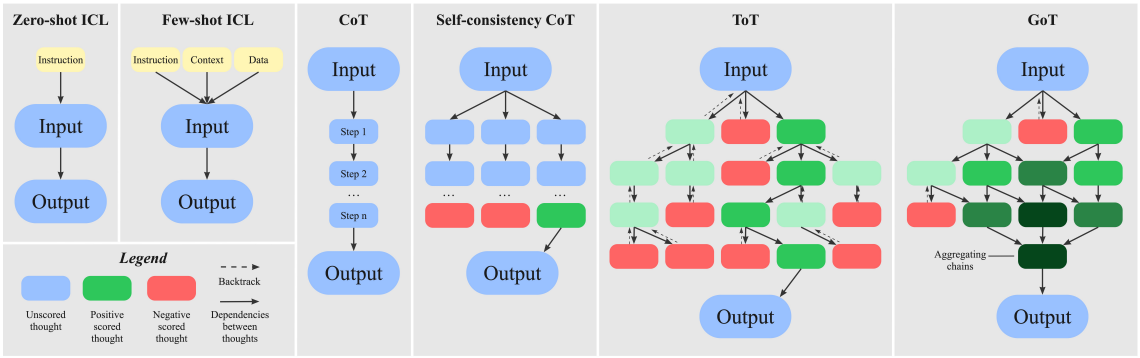


Figure 1: A selection of prompting strategies in LLMs, from basic zero-shot and few-shot In-Context Learning (ICL) to structured reasoning methods such as Chain-of-Thought (CoT), Self-consistency CoT, Tree-of-Thought (ToT) and Graph-of-Thought (GoT). Adapted from [35, 47, 48].

2.2.3 Limitations of LLMs and Prompt Engineering

While large language models (LLMs) have shown impressive performance across a wide range of language tasks, they also present important limitations that affect their reliability, particularly in real-world or high-stakes applications. These limitations manifest in several key areas: hallucination and factual inconsistency, lack of memory and planning capabilities, and the brittleness of prompt-based control mechanisms.

Hallucinations and Opaqueness
A well-documented issue with LLMs is their tendency to "hallucinate", that is, to generate text that appears fluent and coherent but is factually incorrect, misleading, or self-contradictory. Researchers typically classify hallucinations into three types: input-conflicting, context-conflicting, and fact-conflicting [49, 50]. These errors occur because LLMs lack direct access to structured knowledge bases or external verification tools during inference processing. Instead, they rely solely on learned patterns from massive training datasets, which may contain errors or inconsistencies [51, 52]. Beyond factual accuracy, another major concern is the opaque nature of how LLMs arrive at their outputs. These models operate as black boxes: their internal logic is shaped by billions of parameters and complex interactions, making their reasoning difficult to interpret [53]. This lack of transparency raises concerns in fields like healthcare or law, where explainability is essential for trust and accountability [35, 54].

Limits in Planning and Long-Term Consistency
Although LLMs demonstrate generalization through few-shot or in-context learning, they are limited by their fixed context window and lack of persistent memory. They cannot retain information across sessions or reason over extended timeframes. As a result, they struggle with tasks that require multi-step reasoning, long-term planning, or consistent role-based behavior [51]. To address this, recent developments have explored architectures that wrap LLMs in more structured systems. For example, generative agents simulate memory, planning, and behavior consistency by maintaining task history and orchestrating reasoning across time [55]. These approaches reveal that without external scaffolding, LLMs alone are insufficient for tasks requiring structured cognitive processes.

Bias and Societal Risks

Another pervasive challenge is the presence of bias in LLM outputs. Since these models are trained on vast, largely uncurated datasets sourced from the internet, they inherently absorb and sometimes amplify the cultural, social, and ideological biases embedded within those corpora. These biases can manifest in outputs as stereotyping, discriminatory language, or unequal performance across demographic groups [35, 40]. For example, models may generate content that favors dominant cultural narratives, marginalizes certain groups, or produces inappropriate or harmful language in response to benign queries. Such risks are not limited to overt toxicity, but include subtler forms of representational bias that can distort information or perpetuate inequalities. Recent survey literature emphasizes that mitigating bias is particularly difficult due to the opacity and scale of LLMs' training data, and that bias mitigation techniques, such as data filtering, re-weighting, or alignment with human preferences, are only partially effective [40, 36]. This raises important ethical considerations for deploying LLMs in sensitive or regulated domains, where biased outputs may have real-world social, legal, or reputational consequences.

Privacy, Security, and Environmental Impact

Beyond issues of factual reliability and bias, LLMs present additional limitations that affect their suitability for enterprise or high-stakes environments. One concern is privacy leakage: because training corpora often include inadvertently captured personal or confidential data, there is a non-trivial risk that LLMs may memorize and regurgitate sensitive information during inference [36, 51, 40]. Adversarial prompt attacks can exploit such memorization, posing a significant security risk for organizations handling private or regulated data. Another emerging concern is the substantial computational and environmental footprint of training and deploying state-of-the-art LLMs. The hardware, energy, and infrastructure required for both development and inference present barriers to accessibility and raise sustainability questions for widespread adoption [51, 35, 36].

Generalization and Domain Adaptation

While LLMs excel at a broad range of generic language tasks, they frequently under perform when faced with domain-specific problems requiring expert knowledge, up-to-date information, or nuanced understanding of specialized contexts [56, 35, 36]. Without targeted fine-tuning or carefully constructed prompts, model outputs may lack the depth, accuracy, or reliability needed for professional applications in law, medicine, or scientific research. Moreover, LLMs' performance can degrade significantly in low-resource languages or non-standard dialects, reflecting the imbalanced representation of languages

and communities in their training data [51, 36].

Accountability and Explainability

The complexity and black-box nature of LLMs make it difficult to provide clear explanations for specific outputs or errors. This lack of interpretability complicates efforts to establish accountability, particularly in domains where automated decisions or recommendations must be justified to users, regulators, or other stakeholders [56, 36]. As a result, deploying LLMs in critical enterprise or public-sector applications requires not only technical safeguards but also organizational policies and oversight mechanisms to address these explainability and accountability gaps.

Prompt Brittleness and Evaluation Challenges

Prompt engineering has emerged as a central technique for guiding LLM behavior. However, this method is often brittle. Small changes in how a prompt is phrased or structured can lead to large differences in output, even when the underlying task remains the same [57, 56]. This unpredictability complicates reproducibility and hinders reliable use across different contexts or domains. The issue is especially pronounced in reasoning tasks. Even slight modifications can interrupt reasoning chains, causing the model to abandon otherwise "correct" lines of thought [51]. Evaluation methods introduce additional challenges. There is currently no universally accepted framework for evaluating large language models. Existing evaluation practices are fragmented across tasks and domains, with many relying on narrow benchmarks or subjective human assessments that vary in consistency and interpretability [58]. These differences complicate direct comparison across models, particularly in open-ended or generative tasks. Efforts to improve prompting, such as ensemble prompting, self-consistency techniques, or automated prompt generation, offer some improvements, but they do not eliminate the underlying fragility of the method [51].

Toward Agentic Systems

Taken together, these limitations suggest that prompting alone cannot support more advanced cognitive capabilities like planning, memory integration, and self-evaluation. This has led to a growing interest in *agentic systems*, where LLMs function as components within larger, modular workflows that manage task decomposition, tool usage, and memory retrieval in a structured way [55, 35]. These emerging approaches offer a path toward more transparent, reliable, and scalable AI systems, moving beyond the limitations of standalone prompting toward models that can operate more like robust, interactive agents.

2.3 Agentic Systems

2.3.1 Conceptual Origins of Cognitive Systems Research

Agentic systems represent a contemporary evolution of a long-standing research ambition within AI: creating computational models that mimic human cognitive processes. Historically, cognitive architectures such as SOAR and ACT-R have systematically attempted to emulate intelligent behaviors by integrating symbolic reasoning, task decomposition, planning, and robust memory systems into structured, modular frameworks [19, 20, 18]. SOAR is a general cognitive architecture developed to model human problem-solving and learning, designed to unify various cognitive functions in a single system through production rules and episodic memory. ACT-R (Adaptive Control of Thought-Rational) is a modular framework that models human cognition by simulating how different mental modules (such as memory, perception, and motor actions) interact to produce intelligent behavior. These frameworks aim to provide transparency and interpretability, enabling clear comprehension and modification of cognitive processes. Each cognitive module, such as memory, perception, or planning, operates independently yet integrates cohesively within a broader architecture, facilitating comprehensive cognitive functionality such as reasoning, learning, and adaptive decision-making [18].

Cognitive architectures are deeply rooted in interdisciplinary fields including psychology, cognitive science, and artificial intelligence. The original intent behind these architectures was to replicate human-like cognitive robustness and adaptability, characterized by proactive planning, memory integration, reflective self-evaluation, and goal-directed actions [18]. These attributes remain central to modern research in agentic workflows that embed language models within modular structures to support memory, reasoning, and cooperative problem-solving capabilities [22, 23].

Despite their conceptual ambition, early cognitive architectures like SOAR and ACT-R, as well as other Good Old-Fashioned AI (GOFAI) systems, faced significant limitations that hindered their real-world applicability. These systems relied heavily on symbolic representations, central control, and clearly defined modules for perception, reasoning, and action [21, 59]. However, this decomposition presupposed that intelligence could be reduced to interacting modules with clean interfaces, a view that was increasingly challenged by evidence from cognitive science and robotics. Such architectures often struggled with brittle behavior, poor adaptability to complex, noisy environments, and

a reliance on human-designed abstractions that failed to capture the situated, dynamic nature of real-world cognition [60, 21].

A major critique was that these systems required extensive hand-engineered knowledge and explicit world models, which made them inflexible and difficult to scale. It was later demonstrated that intelligence could emerge from parallel, decentralized processes directly coupled to the environment, bypassing the need for central representations and offering greater robustness in real-world applications [21]. Other research emphasized the importance of "situated cognition," highlighting how knowledge is enacted in context rather than stored and manipulated as abstract symbols [60].

This critique led to a shift away from monolithic, centrally planned architectures toward more embodied, distributed, and adaptive models. Nonetheless, early neural and connectionist models also had limitations in memory, planning, and generalization, making them insufficient as standalone agentic systems [59, 61]. Only recently have advances in large-scale machine learning, computational resources, and architectural innovations enabled a practical synthesis of symbolic reasoning, flexible memory, and adaptive perception, laying the groundwork for contemporary agentic systems [61].

The current revival of agentic architectures is therefore made possible by: (1) access to vast computational resources, (2) the emergence of large language models with broad generalization capabilities, and (3) new frameworks for orchestrating modular, interactive components at scale [61, 59]. These advances allow modern agentic systems to overcome many limitations of both GOFAI and early connectionist approaches, achieving new levels of autonomy, adaptability, and transparency.

2.3.2 Defining Agentic Systems

Modern agentic systems extend the ambitions of classical cognitive architectures by integrating LLMs into modular, orchestrated workflows that enable autonomous planning, reasoning, tool use, and memory management. Rather than relying on static prompting or isolated prediction tasks, agentic systems organize AI capabilities into discrete, interacting components, such as planners, memory modules, and tool-using agents, that collaborate to complete complex, multi-step tasks [62, 63, 64].

An agentic system can be defined as an AI architecture in which an LLM is embedded within a broader control structure capable of interpreting user goals, planning sequences of actions, interacting with external systems, and managing intermediate outputs through persistent memory. These systems are frequently implemented through Directed Acyclic Graph (DAG) architectures or flow-based orchestration frameworks, which ensure transparency and reusability by making each reasoning step explicit and modular [22, 55].

This design approach marks a return to foundational definitions of autonomous agents in classical AI. Franklin and Graesser's [17] widely cited definition states that "an autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" [17, p. 25]. LLM-based agentic systems satisfy these criteria in digital environments: they perceive through input parsing and retrieval tools, act via API calls or tool invocations, and update internal memory states in response to changing goals or contexts.

What distinguishes modern agentic systems is their hybrid nature: while grounded in statistical language prediction, they exhibit symbolic behavior through explicit planning and structured control flow. Recent surveys emphasize key features of these systems, including reflection loops, tool integration, collaborative multi-agent coordination, and the capacity for dynamic adaptation across heterogeneous tasks [65, 23]. These properties make agentic systems particularly well-suited for knowledge work domains that require transparency, adaptability, and sequential reasoning beyond what standalone LLMs can offer.

2.3.3 Implications to Knowledge Work

The emergence of agentic systems represents a logical progression in the ongoing transformation of knowledge work. Earlier waves of digital augmentation provided tools that enhanced individual capabilities, streamlining access to information, automating trivial tasks, and improving documentation. Agentic systems, by contrast, introduce a more systemic shift: they embed generative AI into structured workflows capable of autonomous planning, memory integration, and tool-based execution. This allows them to function not just as enhancements to individual performance but is envisioned as collaborative partners within distributed knowledge processes.

Recent empirical studies demonstrate that such systems can significantly boost productivity and quality, particularly for less experienced workers, by acting as real-time knowledge scaffolds and feedback generators [26]. Beyond individual augmentation, however, their most profound impact may lie in how they reconfigure team cognition. When embedded into workstreams, agentic systems can serve as persistent, context-aware collaborators, contributing to planning, information management, and decision support. In this sense, they begin to function as participants in hybrid teams: sociotechnical assemblages in which humans and AI agents jointly maintain a shared understanding of goals, roles, and knowledge assets [66].

This collaborative dynamic mirrors what organizational psychology refers to as a Transactive Memory System (TMS): a structure through which group members delegate memory and expertise, trusting each other (or the system) to retrieve relevant information when needed. Agentic systems expand this concept by bringing precision, recall, and scalability to the management of shared knowledge [66]. In doing so, they do not merely automate individual tasks, but begin to redefine the architecture of cognition at the team level, reshaping how work is planned, remembered, and executed across increasingly hybrid human-AI collaboration ecosystems.

2.4 Service Design in Organizational Innovation

Service design is a human-centered approach to shaping services, grounded in principles of co-creation, iteration, and systems thinking. Rather than focusing on isolated tools or deliverables, it looks at how people, processes, and technologies interact across the

full service journey, from frontstage interfaces to backstage operations. Definitions vary slightly across the literature, but a common thread is its emphasis on designing with users, not just for them. [28, 67] This framing helps organizations make sense of complex environments and uncover unmet needs that aren't always visible through traditional analysis.

In internal settings, service design is just as relevant. Employees become the users, and things like tools, workflows, documentation, and cross-functional collaboration become part of the service experience. By visualizing how internal services actually play out, using tools like service blueprints or journey maps, organizations can identify misalignments, break silos, and design more supportive work systems. It shifts the lens from isolated features to how people experience their work holistically, across touch points, departments, and systems. [68] Co-design is a key part of this process. Rather than relying solely on top-down requirements, it brings stakeholders, including front-line staff, into the design process as experts of their own experience. Research shows this approach improves both the process and outcome: it tends to generate more original ideas, improve fit with user needs, and strengthen engagement throughout the project. [69, 70, 71] It also builds organizational capability for future innovation by fostering collaboration and shared ownership.

This human-centered framing is especially relevant for the AI-driven workflow transformation discussed in the previous sections. Agentic AI systems can automate information retrieval, reasoning, and tool use, but without a service lens, they risk becoming isolated technical components. Service design provides the connective tissue: journey mapping, service blueprinting, and iterative prototyping help make AI's invisible decision flows tangible and align them with actual work practices. Co-design techniques further reduce adoption risk by turning affected knowledge workers into active partners, surfacing tacit needs, reframing problems, and validating early concepts. Studies show this not only improves outcomes, but also accelerates delivery and supports long-term adoption. [70, 69]

In this study, service design serves not only as a theoretical lens but as a guiding logic for how the research was conducted. The following chapter outlines how a user-centered, iterative approach was applied to understand real workflows, identify opportunity areas, and prototype AI interventions in close collaboration with the primary users.

3 Method

This chapter presents the methodological approach used to investigate how generative AI can augment knowledge work within a large enterprise context. The study is based on a qualitative, exploratory single-case design [72] and follows a human-centered, iterative research process grounded in service design principles and structured using the ISO 9241-210 standard [73]. This combination enabled both contextual sensitivity and procedural rigor, ensuring that findings were grounded in real-world practice and translated into actionable insights. The research design integrates four methodological strands: data collection, analysis and clustering, opportunity evaluation, and prototype development. These activities were carried out within a clearly scoped organizational setting, in close collaboration with internal stakeholders and with continuous involvement from the primary user group.

Objective 1	To identify recurring workflow pain points and user needs in knowledge work that may be addressable through generative AI intervention
Objective 2	To evaluate and prioritize opportunity areas for generative AI intervention in knowledge work
Objective 3	To design and develop an advanced prototype that demonstrates how generative AI can be meaningfully embedded into a selected knowledge work process

Each objective maps directly to one or more phases in the human-centered design process, as outlined in ISO 9241-210 standard [73]. Table 3 presents the research phases, their purpose, and their alignment with both the study objectives and ISO's design process model.

Table 3: Overview of the research design phases and their alignment with ISO 9241-210 and the study's objectives

Method Phase	**Description**	**Linked Objective**	**ISO 9241-210 Phase**
Data Collection	Contextual immersion and semi-structured interviews to understand current workflows and surface user needs and pain points	Objective 1	Phase 1
Analysis and Clustering	Thematic coding of pain points and clustering into design-relevant user needs and opportunity areas	Objective 1	Phase 2
Opportunity Evaluation	Prioritization of opportunity areas to select a high-value, feasible use case	Objective 2	Transition between Phase 2-3
Prototype Development	Iterative design validation and development of a advanced prototype demonstrating contextual feasibility	Objective 3	Combination of Phase 3–4

3.1 Research Design

The following section outlines the overall structure of the research design, including the empirical context in which the study was conducted and the methodological frameworks that guided its implementation. It is divided into two parts: the case context and user group, and the methodological framework and design logic.

3.1.1 Case Context

This study adopts a qualitative, exploratory single-case design [72], situated within a global enterprise context undergoing early-stage experimentation with generative AI. The objective of the study is not to produce broadly generalizable results, but rather to generate situated insights and actionable design knowledge [27] on how generative AI can augment knowledge workflows in a real-world context.

The study was carried out over a five-month period in close collaboration with the organization's PM function. During this time, the work was conducted on-site, enabling continuous access to team activities, internal artifacts, and live workflows. This setup allowed for ongoing engagement with key stakeholders and direct exposure to the organizational context in which the studied practices occurred. The work was sponsored by a senior stakeholder, hereafter referred to as T, who holds strategic responsibility for process improvement and capability development within the PM function.

The organization follows the Scaled Agile Framework (SAFe) [74], a widely adopted model for applying agile principles at enterprise scale. SAFe structures work around cross-functional agile release trains, aligns team-level execution with strategic themes, and emphasizes a dual-operating model that balances agility with governance [75, 74]. Within this model, the PM function plays a central role in defining features, managing the product backlog, and aligning development with customer and business needs [76]. This structured yet flexible way of working helped surface repeatable workflows and decision points suitable for analysis and intervention.

From the organization's perspective, the goal was to gain tangible insight into how generative AI could meaningfully support professionals across the PM function. These individuals formed the *primary users* in the study, selected because of the cognitively intensive nature of their work.

This aligns with established definitions of knowledge work, which involve activities such as framing problems, synthesizing diverse inputs, and making judgment-based decisions under conditions of ambiguity [24, 29, 13]. These characteristics closely reflect the responsibilities of modern product teams, particularly in large organizations operating within fast-changing and information-dense environments. As such, the PM function represents a strategically relevant context for investigating the potential of generative-AI augmentation.

3.1.2 Methodological Framework and Design Logic

This study is grounded in a service design perspective, which frames the problem space in terms of user experience, contextual value creation, and systemic alignment, rather than isolated tool implementation [28, 67]. Service design emphasizes co-creation, iteration, and organizational embedding across both frontstage and backstage aspects of service delivery [69, 68, 71]. In this framing, generative AI is not treated as a standalone system, but as one component in a broader human–AI collaboration ecosystem. This perspective shaped the study's approach to need-finding, opportunity framing, and the continuous involvement of the primary users throughout the process.

To provide a rigorous structure for this process, the project adopted the ISO 9241-210 standard for human-centred design of interactive systems [73]. This international framework defines a four-phase iterative process: (1) understanding and specifying the context of use, (2) specifying the user requirements, (3) producing design solutions, and (4) evaluating the design against requirements. While originally rooted in ergonomics and usability engineering, ISO 9241-210 is now broadly used to guide user-centered development in both research and industry settings [27].

Although ISO 9241-210 and service design emerged from different disciplinary traditions, their underlying logic is closely aligned. Both follow a human-centered, iterative design process and promote early and continuous user involvement, contextual understanding, and systems-level thinking [28, 71, 70]. In this study, ISO provided the structural backbone for organizing research activities, while service design offered the terminology, methods, and experiential tools that brought each phase to life. This hybrid approach ensured procedural rigor while remaining grounded in the lived realities of the primary user group.

Each phase of the research corresponded to an ISO 9241-210 phase, but was operationalized using service design practices to ensure relevance, participation, and contextual grounding.

Phase 1: Data Collection
This phase addressed Objective 1 of the study and corresponds to Phase 1 of ISO 9241-210, which emphasizes understanding and specifying the context of use. The work began with deep immersion in the organizational setting, commonly referred to in service design as service discovery or immersion, involving techniques such as service safaris and informal ethnographic engagement [67, 68]. It overlaps with what user-centered design literature terms contextual inquiry, a method for understanding user needs through direct interaction with people and systems in their real environment [77]. Activities included participation in on-boarding and training on internal platforms, informal conversations with team members, and active involvement in digital communication channels. These methods reflect service design's emphasis on grasping the lived context of service delivery and internal user experience [68]. To surface workflow pain points and inform the activity coding framework, semi-structured interviews were conducted with the primary user group [78, 79]. These qualitative techniques are foundational in both service design and user-centered systems design [28, 68, 27]. The resulting insights were mapped against a custom two-level activity coding framework, derived from the Scaled Agile Framework and enriched by real-world observations, enabling traceable analysis at both the responsibility area and specific activity levels [28].

Phase 2: Thematic Analysis and Opportunity Clustering
Aligned with Objective 1 and ISO 9241-210 Phase 2, this phase involved thematic coding of the collected data to identify recurring pain points and user needs. In line with service design's commitment to visualizing complexity and aligning needs with value creation [28, 67], findings were synthesized through thematic analysis. Opportunity areas were mapped and clustered, aligning with service design's emphasis on making hidden needs visible and actionable [69, 71].

Phase 3: Opportunity Evaluation and Prioritization
This phase addressed Objective 2 of the study and represents the transition between Phases 2 and 3 in ISO 9241-210. It involved the prioritization of opportunity areas to identify a high-value, feasible use case for generative AI intervention. This was accomplished through a multi-criteria decision analysis (MCDA) framework, which evaluated shortlisted opportunities on dimensions such as business impact, technical

feasibility, risk, and scalability [80, 81]. Evaluation integrated analytic reasoning from coded evidence with participatory engagement, including stakeholder feedback sessions and informal validation conversations [28, 70]. This participatory and iterative approach exemplifies service design's commitment to co-creation and shared decision-making [71], and guided the transition from user requirements to solution development.

Phase 4: Persona Creation, Prototype Development and Iterative Validation
This final phase supported Objective 3 and aligns with Phases 3 and 4 of ISO 9241-210, focused on design solutions and evaluation against requirements. At the outset of prototype development, a user persona was created to represent the intended user group, a smaller set of internal users directly affected by the prioritized opportunity area [68, 27]. The persona served as a design artifact for aligning the development process with the real cognitive and practical environments of these users [69, 71, 70]. The process then centered on the iterative design and development of an advanced prototype, demonstrating how generative AI could be embedded in knowledge work processes. Following service design principles, rapid prototyping and low-fidelity UI concepts were used to explore interaction concepts and make abstract system logic tangible [68, 69]. The most promising opportunity was realized in first an initial high-fidelity prototype, developed using Lovable [82], a low-code platform for rapid prototyping and UI development. Then it was further refined into the final advanced prototype with workflows orchestrated in LangGraph [83], an open-source framework for building stateful, multi-agent applications with LLMs. This combination allowed the prototype to closely mirror real organizational practices and experiment with advanced agentic AI capabilities. Stakeholder feedback was continually integrated through evaluation sessions and artifact reviews, ensuring the prototype responded to actual user needs and contextual constraints [71, 70]. This phase validated both the feasibility and organizational fit of the proposed solution.

Throughout all phases, the process was grounded in service design's core logic: prioritizing co-creation, embracing iteration, and embedding design activities in real organizational contexts. This ensured that design decisions reflected both user needs and system-level constraints, and that the research contributed to both actionable outcomes and broader organizational learning [28, 69, 71].

3.2 Data Collection

The data collection phase aimed to build a situated understanding of current workflows and user needs within the PM function. This was achieved through organizational immersion, informal stakeholder engagement, and semi-structured interviews, following a service discovery approach grounded in user-centered and participatory design principles.

3.2.1 Organizational Immersion and Service Discovery

Access to the organization was enabled through collaboration with T, a senior stakeholder in the PM function, who served as both sponsor and domain expert. This partnership was pivotal for framing the research scope, facilitating early stakeholder engagement, and supporting participant selection.

Before the formal interview phase, an extended exploratory effort was undertaken to build understanding of the organizational landscape, AI maturity, strategic priorities, and technical constraints. This aligns with established service design practices of service exploration and immersion, which stress the importance of research activities embedded in real work settings and across both frontstage and backstage layers of the service environment [67, 68]. The process included digital and in-person meetings with strategic stakeholders from the PM function at various levels, introductory sessions with AI and data science leadership, and additional dialogue with colleagues involved in generative AI-related initiatives. Such immersion supports the surfacing of domain language, recurring pain points, and potential opportunity areas, and is further reinforced by the completion of approximately 20 product training modules via the internal learning platform to gain familiarity with the organization´s portfolio and terminology.

In keeping with a distributed, digital-first working culture, the service discovery phase drew on informal interactions, stakeholder meetings, and exploratory discussions held over Microsoft Teams and in-person channels. Service design literature positions this approach as essential for generating relevant and actionable insight, since it grounds findings in authentic user experiences rather than in abstract or decontextualized observations [68, 70]. Emphasis is placed on engaging with users in context and meeting them within their real-world flow of work, thereby surfacing needs and practices that are not always evident in interviews or surveys alone [67]. The adoption of digital channels

further reduced barriers to participation and enabled broader engagement, which has been recognized as especially valuable in distributed service design settings [54]. This approach facilitated later phases of the research by enabling flexible scheduling, repeated touchpoints, and inclusive involvement throughout the discovery process.

3.2.2 Participant Selection and Sampling Strategy

Participant selection followed a purposeful sampling approach [84, 85], targeting members of the product-management (PM) function who were identified as the intended end-users of the prospective solution. This strategy is widely recognized in user-centered and service design research for enabling engagement with users whose experience and responsibilities are directly aligned with the system or service being designed [27, 70].

An organizational map of the PM function was created in Miro [86], a collaborative digital whiteboarding tool, using internal directory data to visualize reporting lines, team structures, and Business Unit (BU) affiliations. This map served both as a planning tool for study coordination and as a practical artifact for tracking outreach, responses, and interview logistics. All individuals in the PM function across the three BUs, pseudonymously referred to as Ada, Turing, and Hopper, were invited to participate. Out of 20 invitations, 17 interviews were completed with professionals from the PM function distributed across these BUs. This approach ensured that the sample reflected the full spectrum of the PM domain and captured the diversity of career stages, team sizes, and workflow specializations present within the function.

The selection criteria were explicitly aligned with the study's research objectives, with the intention to engage those most likely to be directly affected by, and to benefit from, the generative AI intervention. Purposeful sampling is widely recognized in qualitative research and service design for supporting the collection of rich, practice-based insights from those most closely affected by the research focus [70, 27, 84, 85].

3.2.3 Semi-Structured Interviews

In line with service design's participatory and user-centred ethos, the interview guide was initially framed as a Voice of the Customer (VoC) instrument. VoC is widely used in both service design and product development to capture user needs, expectations, and

pain points in a systematic and actionable manner [87, 28, 71]. The guide was developed based on insights gathered during the service discovery phase and structured around three main thematic areas: current workflows and tool use, pain points and time allocation, and perceived potential for AI-driven support. The complete, de-identified interview guide is provided in Appendix A for reference.

The interviews were conducted in collaboration with another master's thesis team working in parallel on a related project focused on data analytics in the PM function. While the research aims differed, both projects shared an interest in PM practices, allowing for coordinated planning and a more efficient use of participant time. This collaboration contributed to a broader contextual understanding. Each interview session was guided by role-specific objectives, and the generative AI–focused component was introduced and explored independently as part of this study.

Although the full interview guide covered six structured thematic areas with detailed probing questions, the format evolved as familiarity with organizational language and practices deepened. The interviews increasingly converged around four core themes that consistently generated the most relevant insights.

1. Could you summarize your role and responsibilities within product management?
2. What activities do you wish you had more time or support with?
3. What challenges do you face in getting customer and market insights?
4. How do you believe AI could improve product management?

These themes anchored the interviews and enabled a consistent, theory-driven approach, while leaving room for additional probes and follow-up questions to clarify specific workflows or domain processes. This flexible, theme-based structure is recognized in qualitative interviewing and design research as an effective way to balance comparability with responsiveness to the participant's framing [79, 88, 28].

All interviews were conducted remotely via Microsoft Teams, reflecting the organization's digital-first working model. Each session lasted approximately 45–60 minutes. With informed consent, interviews were recorded and automatically transcribed using Microsoft Copilot, enabling rapid and accurate data capture. The resulting rich, contextually grounded material formed the basis for the coding and analysis described in the following section.

3.3 Analytical Framework and Coding

Although service design projects often employ rapid or "quick-and-dirty" analysis methods [28, 71], a more rigorous qualitative approach was warranted in this study to ensure traceability, credibility, and actionable outcomes for both academic and organizational stakeholders. This analytical depth was motivated by the complexity of the enterprise environment, the need to build stakeholder trust in the findings, and the importance of aligning AI intervention opportunities with established workflows and governance structures. Qualitative research literature emphasizes that when the stakes or context demand it, structured coding and multi-stage analysis provide transparency and validity to opportunity identification and solution design [89, 90, 91].

The interview data were analyzed using a three-stage qualitative coding process adapted from grounded theory and qualitative content analysis: open coding, axial coding, and selective coding [89, 90, 91]. This analytical structure enabled a traceable progression from raw transcript content to structured insights and actionable AI opportunity areas. A hybrid approach was employed, combining AI-assisted extraction with manual validation and categorization. This method aligns with ISO 9241-210's emphasis on developing user requirements based on contextual inquiry into work practices and user needs [73], and was tailored to the organizational use of the Scaled Agile Framework (SAFe) to ensure relevance and traceability.

3.3.1 Structuring Interview Content with AI Assistance

Open coding refers to the initial process of breaking down qualitative data into discrete segments, closely examining each part, and assigning conceptual labels to reflect the underlying meaning or function [90, 89]. This stage generates a comprehensive inventory of themes, activities, and pain points expressed by participants, free from predefined categories.

In this study, open coding was conducted using Microsoft Copilot in Teams to accelerate coding and reduce manual overhead to extract structured insights from each interview. A prompt was developed through multiple rounds of testing and iterative refinement. Once finalized, it was used to extract task-level insights and associated metadata without generalization or speculation. The prompt instructed the model to generate detailed,

structured entries for each relevant activity mentioned during the interview. The exact prompt is provided below:

Copilot Prompt Used for Interview Structuring

Title: [Short, descriptive title]
Original Task/Activity: Summarize what the interviewee said about this task
Detailed Description: What the interviewee currently does & any relevant context (stakeholders, frequency, triggers, etc.)
Systems/Tools Involved: List all tools, platforms, or data sources mentioned
Pain Points: What needs or challenges does the interviewee mention?
Potential AI Approach: Were there any specific AI solutions or other solutions discussed related to the task/activity?
Benefits / Value: Time saved, error reduction, improved decision-making, etc.
Constraints / Requirements: Any organizational, technical, and functional/non-functional constraints or requirements for solution development mentioned
Any Additional Notes: Any other relevant insights, comments, or considerations

Instruction: Extract all pain points, manual or time-consuming tasks and needs expressed by the interviewee and organize them into this format. Do not summarize broadly - list every single task, even small ones. Do not fill in information into the categories when there is no mention of it connected to the specific entry.

The output of this prompt was reviewed and validated manually against the full transcript. Hallucinated content, speculative inferences, or auto-filled fields were removed. The AI was re-prompted as needed until saturation was reached, that is, until no new grounded insights could be generated from the transcript. This judgment was based on prior understanding of LLM limitations, particularly their tendency to hallucinate when prompted to provide structured outputs in the absence of explicit source content [50]. This approach significantly reduced the need to reread entire transcripts multiple times, while preserving analytical rigor through iterative, researcher-led review. Each validated entry was then entered in a structured Excel-based dataset, which served as the basis for the next phase of the analysis.

3.3.2 Two-Level Categorization Framework

Axial coding is the process of relating and grouping open codes into higher-order categories or frameworks by identifying connections, hierarchies, or sequences among them [90, 91]. This enables the data to be reorganized around core activities, responsibilities, and recurring workflow patterns.

In the second phase, axial coding was applied by grouping individual entries into a structured categorization scheme. This involved identifying relationships between individual tasks and broader areas of responsibility. A custom two-level coding framework was developed, based on the Scaled Agile Framework (SAFe), which mirrored the formal role definitions and work practices within the organization. The levels were:

Responsibility Area (Level 1): Eight high-level SAFe-aligned product management responsibilities such as Market Research & Analysis, Product Planning & Roadmapping, and Performance Monitoring & Optimization.

Specific Activity (Level 2): Fine-grained task types within each responsibility area, such as "Gather voice of customer data," "Prioritize features using WSJF," or "Monitor feature adoption."

These activity definitions were derived through synthesis of SAFe documentation and enriched with insights gathered through organizational immersion and service discovery. This phase was essential in adapting abstract SAFe roles to actual practices within the organization. The full classification framework is presented in Appendix B. This framework enabled traceable, fine-grained analysis of pain points and opportunity areas within specific responsibility domains.

3.3.3 Thematic Synthesis and Opportunity Clustering

Selective coding is the final phase in which previously categorized data are synthesized around key themes or "core categories," allowing for comparative analysis, aggregation, and prioritization of findings [90, 89]. No new codes are introduced at this stage; instead, the focus is on integrating and refining categories to support interpretation and decision-making.

In the final phase, selective coding was used to synthesize the categorized data into

a structure suitable for comparative analysis and prioritization. No new codes were introduced. Instead, the focus was on preparing the dataset, already categorized by responsibility area and activity, for multi-dimensional aggregation and cross-comparison. Each entry was tagged with consistent metadata, including SAFe responsibility area, specific activity, interviewee ID, and BU affiliation. These structured attributes enabled the dataset to be analyzed using pivot tables and visualized through comparative charts, supporting the identification of workflow concentrations and organizational pain points. These outputs were not used to draw final conclusions, but served to support analytical orientation and prioritization in the next stage. They enabled the identification of areas with concentrated manual effort or recurring pain points, which then could be evaluated through the evaluation framework.

3.3.4 Shortlisting Opportunity Areas

Following selective coding, the synthesized dataset was filtered to identify high-potential opportunity areas for generative AI intervention. Because each entry was already categorized by SAFe-aligned responsibility area and specific activity, the dataset provided a structured foundation for aggregation. Data visualization tools such as pivot tables and distribution plots were used to reveal clusters of pain points and highlight concentrations of recurring user needs. This clustering process directly informed the synthesis of a shortlist of opportunity areas, with each item on the shortlist linked to multiple coded entries and grounded in actual, recurring user needs.

The shortlisting process was grounded in service design principles, including co-creation and iterative refinement [28, 71, 67]. Early insights were shared with internal stakeholders, and informal feedback was solicited to refine the definition and relevance of each opportunity area. This participatory approach helped build a shared understanding of pressing pain points and ensured the shortlist reflected actual practice and organizational context. The resulting shortlist of opportunity areas formed the basis for the structured evaluation and prioritization described in the following section.

3.4 Opportunity Evaluation

The purpose of the opportunity evaluation phase was to prioritize the shortlisted opportunity areas identified during the coding and clustering phases, in order to select a high-value, feasible use case for generative AI intervention. The evaluation process applied a framework which assessed each opportunity according to business impact, technical feasibility, risk, and scalability. Evaluation integrated analytic reasoning derived from the coded dataset with participatory engagement, including feedback sessions and informal validation conversations with key stakeholders. This participatory and iterative approach exemplifies service design's commitment to co-creation and shared decision-making [28, 71, 67], and guided the transition from user needs to prototype development.

3.4.1 Evaluation Framework: Multi-Criteria Decision Analysis

To evaluate the shortlisted opportunities in a structured and transparent manner, a weighted scoring model was employed following established practices in Multi-Criteria Decision Analysis (MCDA) [80, 81]. The goal was not to produce a strictly quantitative ranking, but to facilitate comparative reasoning across diverse opportunity areas. The four evaluation criteria defined were each selected to reflect both business value and technical viability. These criteria are presented in Table 4.

Each opportunity was scored on a qualitative three-point scale (low, medium, high) against each criterion. The scoring process was informed by findings from service discovery, informal stakeholder conversations, and internal knowledge of the organization's technical and governance landscape. Rather than applying rigid technical benchmarks, the evaluation framework was adapted to reflect the lived realities of knowledge workers, an approach aligned with the situated, user-centered perspective of service design [67, 68].

Table 4: MCDA Framework with scoring criteria definitions

Criteria	Rating	Definition
Impact	High	Affects core business outcomes or a large share of product management work
	Medium	Improves efficiency or quality in a localized or incremental way
	Low	Offers limited or peripheral value
Feasibility	High	Can be implemented using current tools and data; minimal integration challenges
	Medium	Some challenges with data access, integration, or change management, but achievable
	Low	Major blockers such as missing data, technical complexity, or inconsistent workflows
Risk	High	High likelihood of failure, compliance/privacy risks, or user resistance
	Medium	Manageable risks; moderate complexity or regulatory issues
	Low	Low risk of failure; well-understood problem space; easily mitigated
Scalability	High	Can be generalized across teams or domains; value increases with scale
	Medium	Applicable in a specific context or team, with moderate extension potential
	Low	Niche use case with limited applicability elsewhere

3.4.2 Prioritization and Final Selection

Once the MCDA scoring was complete, the shortlisted opportunities were compared based on their overall profiles. While the scoring provided analytical guidance, the final selection was based on a broader synthesis that included practical constraints, stakeholder alignment, and project scope. Three qualitative principles were applied to support this decision-making process, presented in Table 5. These principles were applied interpretively alongside MCDA scores to identify the most promising opportunity for prototype development.

Table 5: Qualitative selection principles used to guide final prioritization

Selection Principle	Description
Repetitiveness / Cognitive Load	The task involved frequent repetition or significant cognitive effort, suggesting clear value in automation or augmentation
Data Availability	Internal documentation, legacy research, or public data sources were already accessible, enabling rapid prototyping
Fit with GenAI Capabilities	The task aligned with the known strengths of generative AI, such as text or data generation, summarization, clustering, or information synthesis

To validate the proposed direction, a series of individual meetings were conducted with key stakeholders. These included the prospective end-users, T, and two stakeholders with strategic responsibilities surrounding AI initiatives. In each session, pain point distributions were visualized and discussed using selected charts and pivot tables derived from the coded dataset. The aim was to ensure that the prioritization accurately reflected actual workflow pain points and was seen as both relevant and feasible by stakeholders.

These individual feedback sessions served not only as analytical checkpoints but also as service design validation touchpoints [70]. By presenting early synthesized findings and involving both prospective end-users and strategic stakeholders in interpreting the emerging opportunity areas, the process aligned with key service design principles, particularly co-creation, participatory prioritization, and iterative refinement [70, 27, 69]. Rather than relying solely on top-down evaluation or technical fit, the final decision emerged from a shared understanding of the most meaningful opportunity area and promising solution spaces.

This participatory grounding strengthened the relevance of the final selection and ensured that the prototype would be embedded in actual work practices, not layered on top of them. The selected use case, detailed in the Results chapter, reflects this combination of structured evaluation and service design-oriented logic.

3.5 Prototype Development

Based on the prioritization process the prototype focused on a selected opportunity area related to *Manual and Inefficient Market and Competitor Research*. The aim was not to deliver a production-ready solution, but to explore the feasibility of AI-supported interventions and to assess how such a system could integrate into existing internal knowledge work practices. The development process followed a service design-informed approach, emphasizing co-design, iteration, and close alignment with user context.

3.5.1 Scoping and Stakeholder Involvement

To scope the prototype effectively, a specific prospective end-user group was identified as those with primary responsibility for conducting competitor and market research across the three BUs. Although this activity was often deprioritized, this group consistently expressed a strong need for better support and regularly invested significant efforts into this type of work. The group consisted of four individuals currently responsible for such research, one in Ada, two in Turing, and one in Hopper. An additional stakeholder, who had previously held this responsibility in Hopper but had since moved to another part of the organization, was also included due to their historical knowledge and continued interest in the solution. Their shared responsibilities, workflows, and needs formed a coherent user group suitable for participatory engagement. In service design terms, they were treated as internal service users whose experiences, expectations, and constraints needed to be integrated into the design process [68, 67].

To facilitate situated understanding and frame design decisions around user needs, a persona was developed based on data from previous interviews and informal conversations. In line with participatory design practice, the persona acted as an experience-based abstraction: a representation of recurring goals, pain points, and expectations among the user group [27]. The persona served as a design artifact that made tacit knowledge visible and helped align the development process with the real-life cognitive and practical environments of the intended users [71, 69].

Beyond user profiles, participants were also asked to contribute relevant design artifacts, including examples of past competitor analyses, common research questions they were asked to investigate, and estimates of both the time typically available for such tasks and

the time they actually required. These artifacts helped to ground the prototyping effort in authentic service interactions, consistent with service design's emphasis on contextual insight and system-wide alignment [68].

The prototype was scoped as a lightweight, situated intervention, intended to explore feasibility while minimizing disruption to existing organizational systems and roles. This reflects the service design principle of starting with low-complexity, high-value entry points that can be tested and refined iteratively within local workflows [28]. Rather than pursuing functional completeness, the aim was to assess how generative AI could be meaningfully introduced to support complex, knowledge-intensive research tasks in ways that respected existing routines and decision structures.

To ensure the prototype was both manageable and relevant for early-stage evaluation, the workflow was narrowed to support a representative research task: the structured exploration of publicly available data for competitor analysis. The prototype was explicitly framed as a service touchpoint [68], with the design focusing on how users would interact with the end-to-end research process across digital tools, knowledge artifacts, and established analysis workflows. Evaluation of the prototype prioritized two core criteria: the functional fit of the system for supporting the intended research tasks, and the perceived ease of use, clarity, and alignment with actual user workflows. This dual focus reflects best practice in both service and interaction design, where formative testing emphasizes how well a solution supports the target work and how simple and effective it feels to end users [27, 92].

By narrowing the initial workflow and evaluating both task support and user experience, the prototype could demonstrate end-to-end feasibility while providing actionable feedback for iterative refinement. This approach also aligns with service design's systemic orientation toward future service evolution, reuse across touchpoints, and adaptability over time [71, 27].

3.5.2 Design Exploration and UI Conceptualization

To explore how a generative AI solution could be embedded into existing research practices, a series of user-interaction (UI) concepts were developed and tested at low fidelity. This design phase focused on making abstract service ideas tangible and assessable. This is a central tenet of service design, where prototypes are used not only for testing

functionality but also for enabling shared reflection and reframing [68, 67].

Three distinct UI concepts were created using UXCopilot.ai, each representing a different interaction paradigm; (1) A chat-style interface, mimicking popular generative AI tools and offering open-ended prompts and responses; (2) A wizard-like assistant, guiding users through a series of structured input fields and sequential prompts; (3) A modular dashboard, presenting pre-defined input sections and configurable outputs in a more structured format.

These concepts served not as final UI candidates, but as symbolic artifacts used to visually represent early design assumptions and reflect on interaction models [69]. No formal evaluation was conducted at this stage. Instead, the UI concepts were used internally and in informal discussions. This approach enabled rapid feedback on expectations and interaction flow without the overhead of high-fidelity development. The use of multiple interaction paradigms at this stage also supported collaborative sensemaking [70]. Early representations can act as conversation starters that allow users to reflect on their own processes, surfacing tacit needs that may not emerge through interviews alone. In this case, the visual differences between the UI concepts triggered insights about when and why users preferred structure versus flexibility in their research flow.

3.5.3 Prototyping and Technical Exploration

The high-fidelity prototyping phase focused on exploring the technical feasibility of supporting a scoped research task using generative AI, while ensuring the solution remained grounded in the real-world practices, artifacts, and expectations of internal users. Rather than aiming for engineering completeness or integration into enterprise systems, the intent was to build an advanced prototype that could illustrate the core service logic in action and facilitate reflection on its role within existing workflows.

Initial Prototype
The initial prototype was developed using Lovable [82], an AI-powered platform that enables users to create full-stack web applications by describing their ideas in natural language, eliminating the need for coding expertise. This platform facilitated rapid experimentation with interface concepts and simple LLM-powered flows. For this prototype, OpenAI's GPT models [93] were utilized for language generation, while Firecrawl, a web scraping tool designed to convert websites into LLM-ready data formats

[94], was employed to retrieve external textual content. While this low-code setup allowed for swift iteration, it revealed limitations in backend control, error handling, and workflow customization. These constraints reinforced the need for a more adaptable and observable framework.

Advanced Prototype
In line with service design's emphasis on iterative development in context [27], a more advanced prototype was developed by building on top of the open-source Open Deep Research framework [95] by LangChain AI [96]. This framework served as a conceptual and architectural foundation for orchestrating multi-step research tasks, including query parsing, content retrieval, synthesis, and structured summarization. Rather than using the base framework directly, a custom version was developed to reflect the specific requirements of the scoped research use case. The prototype was adapted to incorporate tailored prompt structures, source filtering logic, and output formatting aligned with internal workflows and user expectations. The setup was complemented by LangGraph UI, a visual interface through which system behavior could be configured and observed interactively [97]. Various research queries were tested, parameters such as scraping depth were adjusted, and the evolution of task logic across different runs was monitored. Through LangGraph UI, greater transparency in task logic was also achieved, and dynamic, agent-based research behaviors were simulated in real time.

As the final component of the advanced prototype, a high-fidelity user interface demo was created in Lovable [82] to visualize how such a system could surface in a user-facing tool. This prototype included both dummy data and sample outputs generated by the implemented flow. Its role was to function as a tangible communication artifact, enabling stakeholders to see how a generative AI–supported workflow might look and feel in context. This visual representation supported internal alignment, made abstract system logic discussable, and provided a foundation for further refinement. In line with service design principles, this representation helped make invisible system logic tangible and discussable, acting as a bridge between backend capability and user experience [68, 69].

3.5.4 Design Principles and UI Rationale

The advanced prototype was designed as a lightweight, web-based research assistant intended for iterative, low-disruption integration into existing PM workflows. In line

with established interface design principles, the UI and interaction model were guided by best practices from visual design and cognitive ergonomics [92]. Key principles included the use of whitespace for clarity and separation, strong contrast for readability, consistent alignment to create visual order, and grid-based structure to support modularity and scalability. The layout and navigation were informed by Gestalt laws of perception, specifically grouping, hierarchy, and figure-ground integration to facilitate rapid information processing and pattern recognition [92]. Color palette, proportion, and visual balance were selected to minimize cognitive strain and support selective attention, drawing on recommendations for simplicity and elegance in user interfaces [92].

These principles were applied throughout the development process, ensuring that each design iteration maximized usability, reduced visual noise, and enabled users to efficiently interpret and act on information in alignment with their workflows.

3.5.5 Demo Presentation and Informal Validation

Once the advanced prototype was finalized, the solution was presented in a live demo session with senior stakeholders, including key decision-makers within the organization. The primary aim was to validate the perceived value, feasibility, and relevance of the proposed system and to secure organizational support for future development. This interactive presentation enabled stakeholders to experience the solution's workflow and interface first-hand, ask questions, and provide immediate qualitative feedback. In parallel, the advanced prototype demo was shared directly with all participating end users via a dedicated communication channel.

This final phase reflects the participatory and iterative ethos of service design, where solutions are not only tested in isolation but also communicated and refined through direct engagement with both users and organizational sponsors [71, 27]. Informal validation of this kind serves to build shared ownership, surface final adjustments, and ensure that future development is grounded in collective endorsement.

3.6 Ethical Considerations

The study was carried out in accordance with established ethical guidelines for research involving human participants, as outlined in the Swedish Codex rules for good research

practice [98]. Ethical considerations focused on informed consent, confidentiality, and secure data handling.

All participants were recruited on a voluntary basis and were informed about the purpose, scope, and intended use of the research. Verbal informed consent to record the interview was obtained before each interview.

Confidentiality was maintained throughout the research process. Although the organizational setting was known to participants and referenced during interviews, all individuals and internal artifacts were treated as confidential. Descriptions of company structures, tools, and roles were generalized to prevent attribution, and no direct quotations or identifiable data were included in the final report.

Interview recordings and notes were stored in a secure and GDPR-compliant manner. Access to raw data was limited to the researcher, and the material was used exclusively for analytical purposes within the academic scope of the thesis.

Transparency was maintained with the host organization during all phases of the research. Ethical principles of openness, accountability, and integrity were upheld throughout, and every effort was made to ensure that participants' perspectives were handled respectfully and responsibly.

4 Results

4.1 Pain Point Analysis and Shortlisted Opportunity Areas

This section presents the structured results derived from the data collection and subsequent coding process. Insights were derived from semi-structured interviews and supporting artifacts, and insights were analyzed and structured using the two-level coding scheme described in the method. Each pain point was assigned to a SAFe-aligned responsibility area and categorized by its associated activity. The analysis reveals which areas of PM contain the highest concentration of pain points and provides quantitative grounding for subsequent shortlisting of opportunity areas.

4.1.1 Overview of Dataset

The analysis is based on 122 documented pain points gathered from the semi-structured interviews with the 17 participants. Each entry was categorized using the two-level coding framework described in the method, combining eight SAFe-aligned responsibility areas with a corresponding set of key activity categories.

Table 6 provides a high-level summary of how pain points were distributed across the three participating BUs. The relatively even distribution, with 47, 35, and 40 entries respectively, supports the generalizability of patterns identified in the subsequent analysis. This even distribution indicates that pain points are not isolated to a single unit or context, but rather reflect challenges experienced throughout the organization.

Table 6: Total number of pain points per business unit

Business Unit	Pain Points
Ada	47
Turing	35
Hopper	40
Total	**122**

Table 7 provides a granular view of the dataset, giving an detailed overview of the distribution of pain points by responsibility area, activity type and BU.

Table 7: Distribution of pain points across responsibility areas and key activities, grouped by business unit

Responsibilities and Key Activities	**Ada**	**Turing**	**Hopper**	**Total**
Other	**8**	**4**	**5**	**17**
Communication and administration	2	3	5	10
Sales & Revenue Operations	6	1		7
Market Research & Analysis	**13**	**4**	**11**	**28**
Conduct market research	2		4	6
Conduct research on regulatory and compliance standards	1		1	2
Gather voice of customer data	2	1	3	6
Identify target market	3		1	4
Monitor market trends and competitive landscape	4	2	2	8
Perform opportunity identification and sizing		1		1
Unsure/Other	1			1
Product Strategy & Vision	**2**	**7**	**2**	**11**
Define product vision and strategy	1	4		5
Develop and communicate value streams		1	1	2
Ensure alignment between product objectives and strategic themes	1	2	1	4
Product Planning & Road Mapping	**1**	**2**	**3**	**6**
Align backlog priorities with business objectives	1	1	1	3
Create and maintain product roadmap			1	1
Prioritize features using WSJF		1	1	2
Product Development & Delivery	**5**	**4**	**2**	**11**
Collaborate with ARTs for feature development	1			1
Manage issue triaging and resolution workflows	4	3	2	9
Track development progress through iteration planning		1		1
Launch Planning & Execution		**1**		**1**
Coordinate cross-functional teams for launches		1		1
Performance Monitoring & Optimization	**15**	**9**	**10**	**34**
Define and track product success metrics	2	1	4	7
Measure feature adoption and usage analytics	4	1	4	9
Monitor cost vs. value impact of decisions	2		1	3
Monitor customer feedback loops	6	5	1	12
Optimize product based on data-driven insights	1	2		3
Product Evolution & Lifecycle Management			**2**	**2**
Manage product end-of-life decisions			2	2
Product Governance & Portfolio Management	**3**	**4**	**5**	**12**
Balance demand vs. capacity across teams		2	1	3
Ensure regulatory compliance and risk management			1	1
Manage financial forecasting for investment	1	1		2
Unsure/Other	2	1	3	6
Total	**47**	**35**	**40**	**122**

While the majority of entries align with formal PM responsibilities, a subset of recurring themes fell outside these predefined categories. These were grouped under "Other" and include pain points related to communication, coordination, and challenges originating in adjacent functions. Such entries reflect the cross-collaborative nature of many roles and more general pain points associated with knowledge work.

A closer examination of the distribution of pain points across Ada, Turing, and Hopper reveals a broadly similar pattern, but with subtle differences that reflect local variations in focus and workflow complexity. All three BUs report the highest concentrations of pain points in Performance Monitoring & Optimization (Ada: 15, Turing: 9, Hopper: 10) and Market Research & Analysis (Ada: 13, Turing: 4, Hopper: 11), indicating that challenges related to tracking product performance and synthesizing market intelligence are pervasive across the organization. However, Ada stands out with a particularly high frequency in Market Research & Analysis, while Turing registers the highest pain points in Product Strategy & Vision (7 compared to 2 in both Ada and Hopper), suggesting a stronger strategic orientation or related bottlenecks in that BU. The "Other" category, which captures cross-functional, administrative, and sales-related challenges, also varies: Ada (8) and Hopper (5) report higher levels than Turing (4), possibly reflecting differences in coordination overhead or organizational structure. Despite these variations, the overall distribution of pain points is relatively even, and most responsibility areas surface across all three business units. This pattern indicates that while certain pain points are more pronounced in specific contexts, the underlying challenges are consistent across the PM function.

The data shows that the highest number of pain points were concentrated in Performance Monitoring & Optimization (34), Market Research & Analysis (28), and the "Other" category (17). This suggests that challenges related to tracking product performance, conducting market research, and fulfilling cross-functional or administrative tasks are particularly prevalent. By contrast, areas such as Launch Planning & Execution and Product Evolution & Lifecycle Management surfaced fewer pain points from the analysis of the interviews.

Only the activities that had at least one pain point coded to them are shown as results. Several responsibility areas include additional activities in the underlying categorization framework, but these did not surface from the analysis of the interviews. A full list of all considered key activities per responsibility area is provided in Appendix B.

4.1.2 Pain Points by Responsibility Category

Figure 2 and Table 8 further illustrate this pattern by ranking responsibility areas by total number of pain points. The majority of pain points were associated with three areas: Performance Monitoring & Optimization (34), Market Research & Analysis (28), and the "Other" category (17). These areas reflect core PM activities such as tracking product usage and collecting customer insights, as well as time-consuming operational or cross-functional tasks that extend beyond formally defined PM responsibilities.

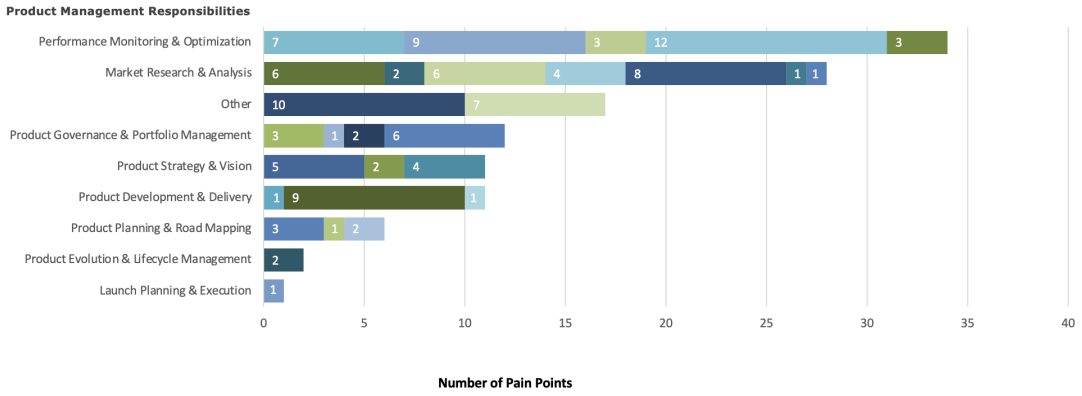


Figure 2: Distribution of pain points across responsibility category. The distinct colors for each column represent the distribution of activities within each responsibility.

Table 8: Responsibility areas ranked by total number of pain points across business units

Rank	Responsibility Area	Ada	Turing	Hopper	Total
1	Performance Monitoring & Optimization	15	9	10	34
2	Market Research & Analysis	13	4	11	28
3	Other	8	4	5	17
4	Product Governance & Portfolio Management	3	4	5	12
5	Product Strategy & Vision	2	7	2	11
5	Product Development & Delivery	5	4	2	11
7	Product Planning & Road Mapping	1	2	3	6
8	Product Evolution & Lifecycle Management			2	2
9	Launch Planning & Execution		1		1
	Total	**47**	**35**	**40**	**122**

4.1.3 Pain Points by Activity Category

Table 9 highlights the ten activities that generated the highest number of pain points and their associated responsibility category. The activities with the greatest number of pain points include monitoring customer feedback, communication and administration, usage analytics, and issue triaging. These clusters point to recurring challenges with insight-gathering, internal coordination, and routine information flows. Notably, activities associated with both customer-facing research and internal operations appear at the top of the list.

Table 9: Top 10 activity categories with the highest number of pain points, ranked by frequency

Rank	Activity Category	Pain Points	Associated Responsibility Category
1	Monitor customer feedback loops	12	Performance Monitoring & Optimization
2	Communication and administration	10	Other
3	Measure feature adoption and usage analytics	9	Performance Monitoring & Optimization
3	Manage issue triaging and resolution workflows	9	Product Development & Delivery
5	Monitor market trends and competitive landscape	8	Market Research & Analysis
6	Sales & Revenue Operations	7	Other
6	Define and track product success metrics	7	Performance Monitoring & Optimization
8	Gather voice of customer data	6	Market Research & Analysis
8	Conduct market research	6	Market Research & Analysis
10	Define product vision and strategy	5	Product Strategy & Vision

The following analysis examines the distribution of pain points within each major responsibility category, combining a pie chart for visual overview and a table for exact counts.

4.1.3.1 Performance Monitoring & Optimization

Figure 3 and Table 10 show that pain points within this area were most densely clustered around customer feedback loops (12), usage analytics (9), and tracking success metrics (7). This suggests that the PM function face persistent challenges in collecting, analyzing, and acting on product performance data.

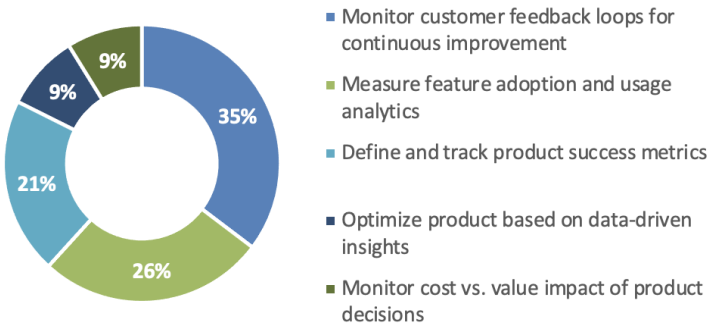


Figure 3: Distribution of pain points across activities within the responsibility category. The highest concentration of pain points is seen in feedback loops, usage analytics, and success metric tracking.

Table 10: Distribution of pain points by activity category

Activity Category	**Pain Points**
Monitor customer feedback loops for continuous improvement	12
Measure feature adoption and usage analytics	9
Define and track product success metrics	7
Optimize product based on data-driven insights	3
Monitor cost vs. value impact of product decisions	3
Total	**34**

4.1.3.2 Market Research & Analysis

Figure 4 and Table 11 reveal that market trend monitoring, customer research, and gathering voice of customer data account for the majority of pain points in this area, underscoring the high manual effort required to synthesize market intelligence.

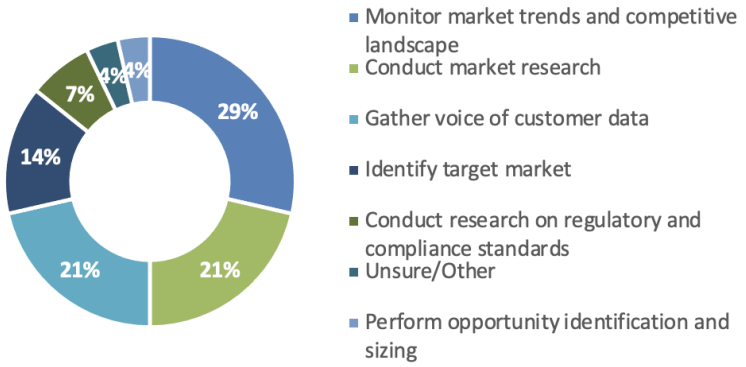


Figure 4: Distribution of pain points across activities within the responsibility category. Pain points were most densely clustered around trend monitoring, customer insight gathering, and general market research activities.

Table 11: Distribution of pain points by activity category

Activity Category	Pain Points
Monitor market trends and competitive landscape	8
Conduct market research	6
Gather voice of customer data	6
Identify target market	4
Conduct research on regulatory and compliance standards	2
Unsure / Other	1
Perform opportunity identification and sizing	1
Total	**28**

4.1.3.3 Product Development & Delivery

Figure 5 and Table 12 indicate that most pain points in this area are concentrated in issue triaging and resolution workflows (9 of 11 total), highlighting coordination and operational bottlenecks.

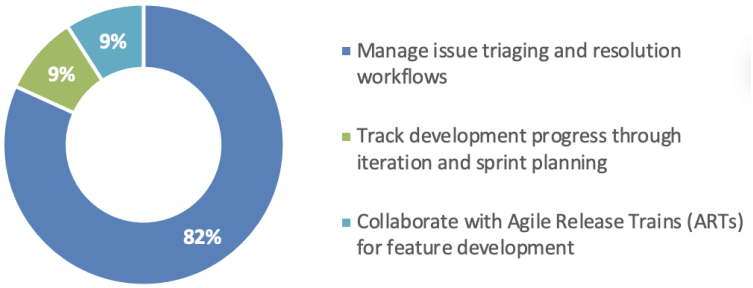


Figure 5: Distribution of pain points across key activities within the responsibility category. Most pain points were clustered around issue triaging and resolution workflows.

Table 12: Distribution of pain points by activity category

Activity Category	**Pain Points**
Manage issue triaging and resolution workflows	9
Track development progress through iteration and sprint planning	1
Collaborate with Agile Release Trains (ARTs) for feature development	1
Total	**11**

4.1.34 Other

Figure 6 and Table 13 show that communication, administration, as well as sales and revenue operations comprise the bulk of pain points coded as “Other.” These findings illustrate the time-consuming nature of responsibilities that fall outside the formal framework but are critical to effective cross-functional work.

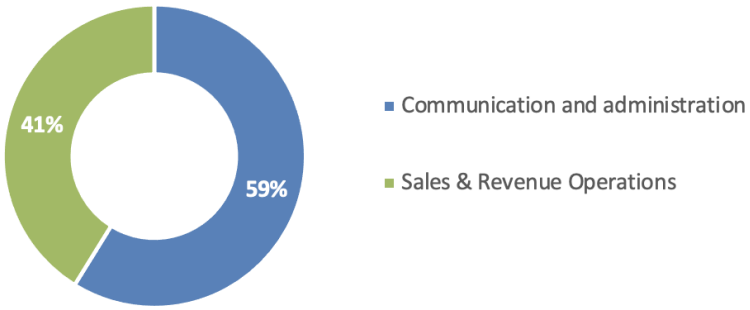


Figure 6: Distribution of pain points across activity types categorized as “Other,” which includes responsibilities outside the formal framework. Clusters here relate to internal coordination, administrative support, sales and revenue operations.

Table 13: Distribution of pain points by activity category

Activity Category	**Pain Points**
Communication and administration	10
Sales & Revenue Operations	7
Total	**17**

4.1.4 Summary of Structured Pain Point Analysis

The structured analysis demonstrates that pain points are highly concentrated in a small number of responsibility and activity categories. While Performance Monitoring & Optimization, Market Research & Analysis, and "Other" emerged as the top responsibility areas, activity-level analysis identified tasks such as customer feedback loops, communication, usage analytics, and issue triaging as particularly pain point-dense. These findings motivated the synthesis of a shortlist of high-impact opportunity areas, which are presented in the following section.

4.1.5 Shortlisted Opportunity Areas

The opportunity areas presented in this section are directly grounded in the pain point mapping and frequency analysis summarized in Table 9, which identified the top activity categories with the highest number of pain points across PM workflows. The shortlist focuses on the activities that, according to both quantitative and qualitative findings, account for the greatest share of inefficiency, manual effort, and organizational risk.

Several points should be noted in how this shortlist was constructed. First, while the majority of opportunity areas correspond closely to the ten most pain point-dense activities, some thematic consolidation was applied for clarity and analytical rigor. Specifically, "Monitor market trends and competitive landscape" and "Conduct market research" were grouped together under a single opportunity area, as the analysis of the interviews and pain point descriptions revealed substantial overlap in both the underlying work and the associated challenges.

It is important to acknowledge that responsibilities for market research and competitor analysis can be distributed across different roles and adjacent commercial or strategy functions. The findings in this case suggest that, in practice, the distinction between competitor analysis and market research is frequently blurred. Interview participants reported that both activities often share data sources, analytical methods, and workflow pain points, and are conducted in close collaboration or even interchangeably. This overlap supports the decision to treat them as a single, holistic opportunity area, capturing the full spectrum of research and insight-generation tasks present in the organization.

Second, the activity "Sales & Revenue Operations" was not prioritized for further opportunity analysis, despite its relatively high pain point count. During synthesis, it was determined that these pain points fell primarily within the domain of revenue operations and support functions, and were therefore outside the main focus on PM work. However, the results and insights relevant to this category were shared with the Revenue Operations team to inform their ongoing process improvements.

Finally, while "Define product vision and strategy" appeared among the top ten pain point categories, these pain points were assessed as representing broad, high-level strategic alignment and leadership challenges. The analysis of the interviews suggested that these issues, while important, are less amenable to targeted process or workflow automation and are more appropriately addressed through organizational and leadership development

initiatives. As such, they were not included in the shortlist of opportunities for generative AI-enabled intervention.

The resulting shortlist therefore reflects a set of opportunity areas where recurring pain points, high manual workload, and thematic clustering suggest strong potential for impactful intervention. Each area is presented below with a concise problem statement and summary of key challenges.

4.2.1.1 Fragmented and Inefficient Management of Customer Feedback Loops

Problem statement: The PM function face ongoing challenges in collecting, aggregating, and analyzing customer feedback from multiple sources, including surveys, support tickets, interviews, and direct outreach. Feedback is frequently fragmented across tools, leading to low response rates, inconsistent categorization, and significant manual effort to synthesize actionable insights. As a result, it is difficult to identify emerging issues, prioritize improvements, and demonstrate the impact of changes to customers in a timely way.

Key challenges:

- Feedback is dispersed across channels and tools, resulting in inconsistent data and missed insights
- Manual aggregation and analysis of qualitative and quantitative feedback is time-consuming and error-prone
- Low customer engagement with surveys and feedback channels reduces the quality of input
- Categorizing and clustering feedback by themes is inconsistent and delays integration with product planning
- Action items are not systematically tracked or revisited

4.2.1.2 Communication and Administration Overhead

Problem statement: The PM function spend a significant portion of their time on internal communication, meeting documentation, email management, and the creation and updating of content such as marketing materials, presentations, and reports. These tasks are highly repetitive, often require coordination across multiple platforms, and are largely manual leading to inefficiency, risk of missed information, and reduced focus on higher-value strategic work.

Key challenges:

- High manual workload for processing, sorting, and responding to large volumes of email and system notifications
- Time-consuming meeting coordination and documentation, including minute-taking and scheduling
- Repetitive creation, updating, and refinement of content across multiple tools
- Difficulty maintaining context, consistency, and accuracy in communication and content over time

4.2.1.3 Limited Visibility into Product Usage, Feature Adoption, and License Analytics

Problem statement: Tracking and analyzing product usage, feature adoption, customer engagement, and license compliance is a highly manual and fragmented process across teams. Data is collected from disparate systems, requiring time-consuming reconciliation, repeated manual reporting, and ad hoc analysis. This leads to delays, errors, incomplete insights, and missed opportunities for proactive action on churn, renewals, or product optimization.

Key challenges:

- Manual reconciliation and reporting of product usage, license entitlements, and customer engagement data across multiple tools and formats
- Limited real-time visibility into customer usage, renewal status, and feature adoption
- High risk of errors and missed revenue due to manual processes, especially for license renewals and compliance
- Difficulty identifying trends and patterns to inform product development, customer retention, and feature prioritization
- Lack of integration and automation prevents holistic and timely decision-making

4.2.1.4 Manual and Fragmented Issue Triaging and Resolution Workflows

Problem statement: The PM function face recurring challenges in efficiently managing and prioritizing a high volume of support tickets, bugs, and customer issues. The current processes are fragmented across multiple tools, involve significant manual review and categorization, and require substantial coordination between support, engineering, and product roles. These workflows are time-consuming, increase the risk of errors or missed information, and make it difficult to ensure timely resolution, particularly for repetitive or high-priority issues.

Key challenges:

- Manual triage and categorization of issues is time-consuming and repetitive
- Difficulties tracking status, ownership, and progress, leading to delayed or missed resolutions
- Lack of integration and automation increases cognitive load for teams
- Repetitive queries and common issues consume significant team resources
- Limited visibility into patterns, root causes, or duplicate/recurrent issues

4.2.1.5 Manual and Inefficient Market and Competitor Research

Problem statement: The PM function spend significant time on manual, repetitive market and competitor research. Insights must be gathered from a wide variety of sources; internal documents, analyst reports, call logs, customer feedback, online searches, and data platforms. Because this work is highly fragmented and largely unstandardized, it leads to duplicated effort, difficulty maintaining up-to-date knowledge, slow identification of market trends, and limited reuse of prior research. Findings are often stored in personal files or siloed formats.

Key challenges:

- Research efforts are duplicated and siloed across teams and individuals
- Analysis and reporting are time-consuming and error-prone
- Data sources and formats are unstructured and inconsistent
- Trends and competitor activities are difficult to monitor and update in real time
- Insights and reports are difficult to reuse or synthesize for future needs

4.2.1.6 Defining and Tracking Product Success Metrics

Problem statement: The PM function invest substantial effort in manually gathering, reconciling, and reporting on key product success metrics such as feature adoption, customer engagement, financial performance, and support costs. Data is scattered across various systems, often leading to inconsistent, delayed, or incomplete reporting. This manual approach hampers timely decision-making, increases the risk of errors, and makes it difficult to align teams around a shared understanding of product performance.

Key challenges:

- Manual aggregation and reconciliation of metrics from multiple, disconnected systems
- Inconsistent definitions and reporting of success metrics
- Time-consuming data processing that delays insights and decision-making
- Lack of real-time access to reliable, actionable product data for stakeholders
- Difficulty integrating financial, usage, and customer data to provide a holistic view of product health

4.2.1.7 Fragmented and Time-Consuming Management of Voice of Customer Insights

Problem statement: Across the PM function, collecting, centralizing, and acting on voice of the customer insights is a recurring challenge. Data is gathered through interviews, feedback sessions, direct outreach, and surveys, but is often scattered across disparate documents, CRM systems, or ad hoc channels. Manual processing, transcript analysis, and profiling are time-consuming and lack structure, leading to inefficiencies, missed opportunities, and a lack of immediacy in responding to customer needs.

Key challenges:

- Insights are fragmented across multiple formats and storage locations, complicating aggregation and access
- Manual synthesis and analysis of qualitative data is time-consuming and often delayed
- Lack of standardized or automated processes for analyzing interview transcripts and survey responses

- Dirty or incomplete data in eg. CRM system reduces the reliability and accuracy of generated insights
- Difficulty integrating VoC findings promptly into product development and market strategy

4.2 Opportunity Evaluation and Prioritization

4.2.1 Multi-Criteria Evaluation and Prioritization

Each shortlisted opportunity area was systematically evaluated using the Multi-Criteria Decision Analysis (MCDA) framework. This involved assessing each area across four dimensions: *Impact*, *Feasibility*, *Risk*, and *Scalability*. The MCDA evaluation (Table 14) reveals several important patterns. Ratings and justifications are grounded in the pain point analysis and service discovery.

Table 14: MCDA of shortlisted opportunity areas

Opportunity Area	**Impact**	**Feasibility**	**Risk**	**Scalability**	**Comments/Justification**
Manual and Inefficient Market and Competitor Research	High	High	Low	High	Central PM task, GenAI/LLM tools mature, low structural barriers
Manual and Fragmented Issue Triaging and Resolution Workflows	Medium	High	Medium	High	Good data, mature tooling, integration straightforward, immediate productivity gains across teams
Fragmented and Time-Consuming Management of Voice of Customer Insights	High	Medium	Medium	High	High potential to streamline VoC analysis and reduce manual transcript work across units; main challenge is integrating unstructured data
Communication and Administration Overhead	Medium	High	Low	High	High feasibility, quick wins, broad reach, but less strategic impact
Fragmented and Inefficient Management of Customer Feedback Loops	High	Low	Medium	Medium	High potential, but data/process fragmentation and lack of governance mean no quick win; foundational work needed first
Limited Visibility into Product Usage, Feature Adoption, and License Analytics	High	Low	High	Medium	Strong value but limited by fragmented systems and data; requires foundational improvements before AI can deliver value
Defining and Tracking Product Success Metrics	Medium	Low	Medium	Medium	As above: metric misalignment, data/process inconsistency are key blockers

The MCDA evaluation in Table 14 highlights several clear patterns for prioritization. “Manual and Inefficient Market and Competitor Research” and “Manual and Fragmented Issue Triaging and Resolution Workflows” emerge as the highest-priority areas for generative AI intervention. Both combine high or medium impact with high feasibility and scalability, supported by mature tooling and robust data availability. Market and competitor research is particularly well-suited for language model-powered information retrieval and synthesis, while issue triaging can leverage established integrations with IT Ticket Software to drive immediate efficiency gains.

“Fragmented and Time-Consuming Management of Voice of Customer Insights” is also highly impactful and scalable, with significant potential to streamline VoC analysis and reduce manual workload. The main challenge in this area lies in integrating unstructured data sources to enable more reliable and actionable insights.

In contrast, opportunities such as “Fragmented and Inefficient Management of Customer Feedback Loops,” “Limited Visibility into Product Usage, Feature Adoption, and License Analytics,” and “Defining and Tracking Product Success Metrics” are deprioritized for immediate prototyping. While their long-term impact could be substantial, current feasibility is low due to fragmented data, lack of standardization, and insufficient process governance. These areas require foundational improvements in data management and organizational workflows before scalable AI solutions can be effectively implemented.

"Communication and Administration Overhead” remains attractive for quick wins because of its high feasibility and broad applicability. However, this area is also the focus of significant investment from major enterprise platforms such as Microsoft 365, which are rapidly introducing AI-powered task automation features. Given this external momentum, developing a custom in-house solution is not recommended, as it would likely duplicate functionality already available on the market and offer limited additional value. Instead, the organization would benefit more from closely tracking advances in vendor solutions and prioritizing adoption and integration of these mature external tools, rather than investing resources in building an internal alternative.

Taken together, these results direct prototyping and pilot efforts toward the opportunity areas with the strongest balance of organizational readiness, impact, and technical viability. Longer-term opportunities are identified as future candidates as foundational capabilities mature.

4.2.2 Final Selection and Rationale

Following the MCDA, a final prioritization was conducted to determine which opportunity area would be taken forward into prototyping. While the evaluation scores provided structured guidance, the selection also considered organizational relevance, stakeholder alignment, and overall fit with the project scope.

Among the top candidates, "Manual and Inefficient Market and Competitor Research" was selected as the most promising opportunity for generative AI intervention. It combined high impact and scalability with strong feasibility and low implementation risk. The task is well-suited to generative AI capabilities, particularly in areas such as automated web retrieval, information synthesis, and structured output generation. Additionally, internal stakeholders confirmed that this activity is time-consuming, repetitive, and strategically important, making it a natural fit for AI-driven augmentation.

"Manual and Fragmented Issue Triaging and Resolution Workflows" was also recognized as a high-potential area, supported by mature tooling and wide applicability across teams. However, due to project scope constraints, this was retained as a secondary candidate for future exploration. The following section presents the resulting design and development process for the Manual and Inefficient Market and Competitor Research.

4.3 Product Development Results

This section presents the main outputs of the prototyping phase, which focused on developing and evaluating an AI-supported solution for market and competitor research. The results below include user-centered persona, exploratory UI concepts, architectural models, and high-fidelity prototypes.

4.3.1 Persona

A detailed persona was developed to encapsulate the needs and challenges faced by those responsible for market and competitor intelligence within the PM function. "Lucas", depicted in Figure 7, is grounded in data from interviews and informal stakeholder meetings and embodies the most prominent needs and challenges identified across the group.

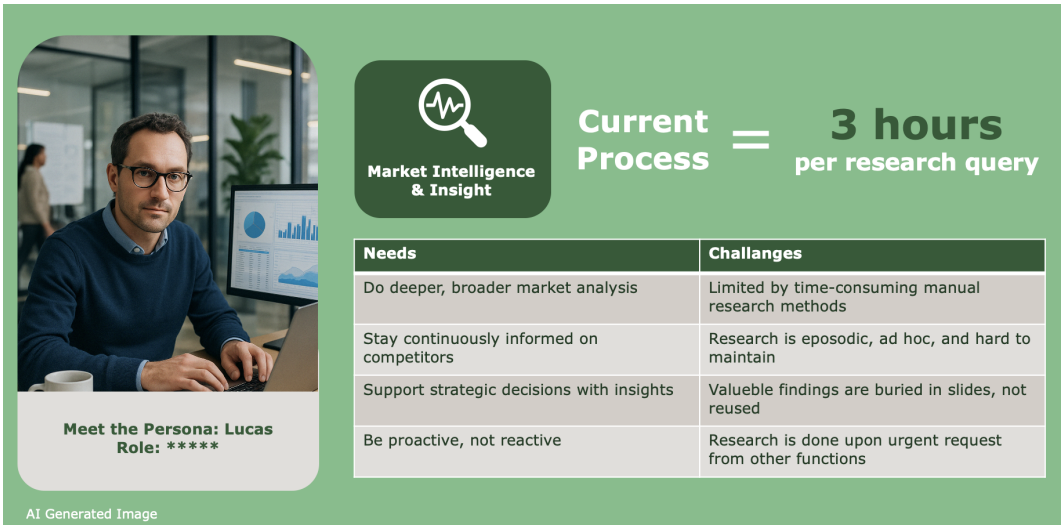


Figure 7: Persona representing the user group involved during the product development phase

The user group behind this persona comprised five individuals, with a gender-balanced composition and an age range of 35–55. Four were currently responsible for market and competitor research across the three BUs, one in Ada, two in Turing, and one in Hopper, while a fifth participant contributed valuable historical perspective and remained engaged as a key stakeholder. Among a wide range of responsibilities, often characterized by dense cross-functional collaboration, this group is specifically tasked with conducting market and competitor research to support strategic and tactical decision-making.

These activities are performed in a context where the time allocated for structured research is highly limited, typically no more than on average four hours per week. As a result, market and competitor research is often conducted episodically, in intensive sessions once or twice per year, or in response to urgent, ad hoc requests from other functions. This persistent time constraint means that deeper analysis and continuous monitoring are frequently deprioritized in favor of more immediate tactical or administrative demands.

The principal needs identified for this group include the ability to conduct deeper, broader market analysis; stay continuously informed about competitors; support strategic decisions with timely insights; and shift from a reactive to a more proactive way of working. However, several persistent challenges impede these objectives. Manual research methods are time-consuming and difficult to scale, making it challenging to

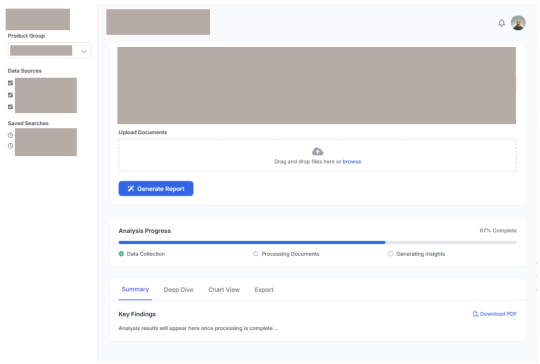
maintain up-to-date intelligence. Research is frequently episodic, ad hoc, and fragmented, which complicates knowledge management and reuse. Valuable findings are often buried in presentations or personal files rather than stored in reusable formats, and urgent requests from other functions regularly interrupt more strategic research efforts.

While the persona illustrated in Figure 7 is male, the actual user group included both men and women in similar roles and with comparable experience levels. The choice to present a single, male persona reflects the synthesis approach used in this project. Future work could benefit from developing a team-based or multi-profile persona to better capture the diversity of experiences and collaboration patterns within the group.

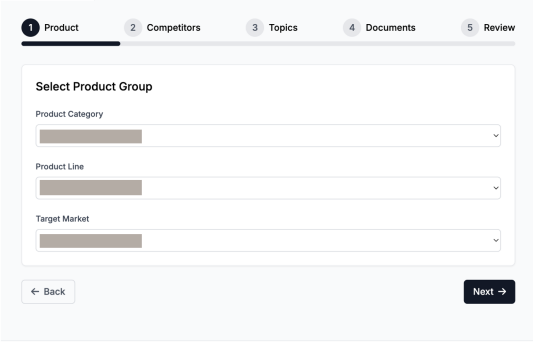
4.3.2 Exploratory UI Concepts

Three alternative UI concepts were created to explore how users might interact with an AI research assistant. These included: (a) A chat-based interface for open-ended prompting, (b) A wizard-style interface for stepwise guidance, and (c) A dashboard model for structured input and output. The concepts, shown in Figure 8, were used to facilitate internal discussion about preferred workflows and interaction patterns.

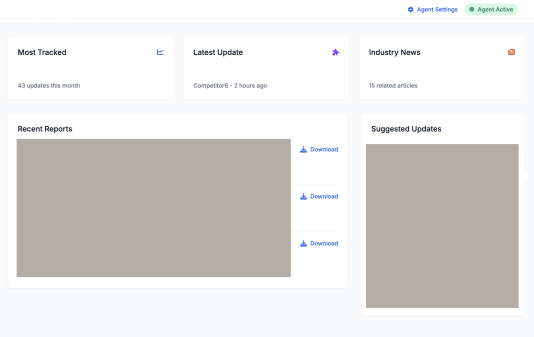
While no formal and systematic usability evaluation was conducted, the informal feedback and visual comparisons revealed important insights. The structured dashboard interface was perceived as the most aligned with how users currently organize and present research tasks. It was described as intuitive, easier to navigate and more consistent. This supported the interpretation that generative AI outputs would need to be clearly structured and reusable, not embedded in freeform dialogue. In contrast, the chat interface was seen as too open-ended and ambiguous for a task that users preferred to approach with a predefined scope and output format. The wizard model, while more guided, was considered overly rigid and did not offer the flexibility required for complex research inputs that vary across tasks.



(a) Chat Interface



(b) Wizard Interface



(c) Dashboard Interface

Figure 8: Exploratory UI concepts

This early-stage concept testing served a dual function: it surfaced user expectations for interaction structure, and it helped reinforce that the generative AI solution would need to integrate seamlessly into existing artifact flows rather than introduce entirely new formats or interfaces. These findings informed the decision to develop the advanced prototype using a dashboard-like interaction logic, with clear input fields and structured outputs.

4.3.3 System Architecture

A conceptual system architecture was developed to illustrate the modular, agentic workflow underpinning the AI-assisted research flow. As shown in Figure 9, the diagram presents a high-level sequential flow reflecting the core logic of agentic systems. User input is parsed and planned, relevant external data is retrieved and cleaned, and results are output in a structured format. The detailed implementation architecture is not disclosed due to confidentiality.

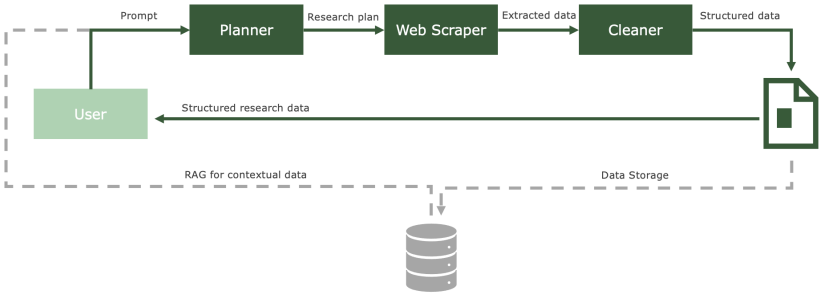
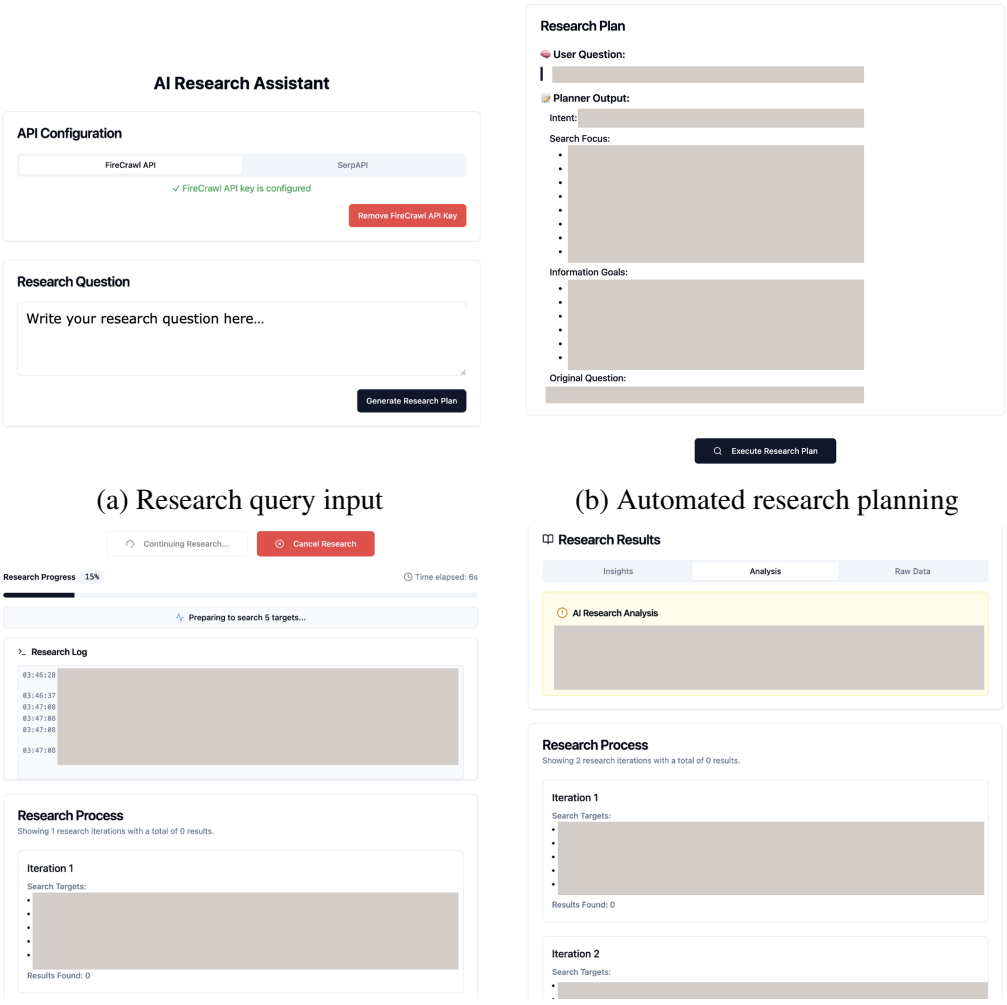


Figure 9: Conceptual system architecture

4.3.4 Initial High-Fidelity Prototype

To concretize and evaluate the envisioned end-to-end research workflow, a high-fidelity prototype was developed. The goal was to make the proposed service logic, user journey, and anticipated output structure tangible, enabling realistic, scenario-based testing and internal stakeholder feedback. This prototype allowed to simulate key interactions, visualize process flows, and surface practical design and integration challenges early in the development process.

Screenshots of the workflow prototype are presented in Figure 10, with each subfigure illustrating a major step in the research flow: Subfigure 10a shows the research query input interface, Subfigure 10b displays the automated research planning, Subfigure 10c presents the web data retrieval and processing step, and Subfigure 10d depicts the structured summary output generated for the user.



(a) Research query input

(b) Automated research planning

(c) Web data retrieval and processing

(d) Structured summary output

Figure 10: High-fidelity workflow prototype

This prototyping exercise surfaced critical limitations with the chosen low-code platform, particularly in terms of backend extensibility and workflow customization. These constraints informed the subsequent decision to pursue a more robust and flexible technical architecture, ensuring that the final solution could accommodate the complexity and adaptability required.

4.3.5 Advanced Prototype

The advanced prototype was developed to explore how generative AI could meaningfully support market and competitor research in a way that respected both organizational constraints and user needs. While detailed descriptions of workflow logic, interface elements, and outputs are withheld for confidentiality reasons, the guiding principles and architectural logic behind the system can be summarized to illustrate its alignment with user needs, responsible AI use, and service design practice.

Human-in-Control and Reliability Safeguards
In response to well-documented limitations of large language models, such as hallucinations, brittle reasoning, and lack of transparency [35, 50, 51], the prototype was explicitly designed to embed human oversight into every stage of interaction. Rather than attempting to automate end-to-end research tasks, the system facilitates a collaborative workflow in which users remain the primary decision-makers. At key points in the process, users are invited to review and validate AI-generated suggestions before continuing. This "human-in-control" orientation follows best practices in human-centered AI design [15, 32] and is essential to ensuring the reliability, traceability, and organizational trust required for high-stakes knowledge work. To reinforce transparency, outputs are annotated with provenance metadata, including timestamps and dynamically generated source references, supporting traceability and future verification.

Design Rationale and UI/UX Considerations
The user interface was developed in line with principles from both cognitive ergonomics and visual interface design [92]. Specific attention was given to whitespace, alignment, and consistent hierarchy, creating an interface that supports rapid interpretation and minimizes cognitive load. The layout reflects Gestalt principles of visual grouping and figure-ground separation, allowing users to focus selectively on relevant content and progress through tasks without distraction. Navigation is streamlined and task-oriented, based on patterns observed in workflow artifacts provided by the user group and reinforced through early prototyping feedback. While the precise color scheme is not disclosed, contrast was optimized for readability and aligned with the organization's brand aesthetic.

Technical Approach and Modular Agentic Architecture
The architecture follows principles of agentic systems as defined in recent research [22, 61, 63], in which language models are orchestrated as modular components within a structured, task-oriented flow. Rather than relying on monolithic prompting, tasks are decomposed into stages, such as parsing, retrieval, synthesis, and formatting, each handled by specialized components. Chain-of-thought prompting and structured intermediate steps were used to scaffold reasoning and increase model robustness [42]. Different model types were orchestrated to balance capability with cost: larger models handled high-complexity reasoning tasks, while smaller models supported lightweight parsing and extraction. These patterns reflect a shift away from static, black-box prompting toward more transparent, interpretable workflows, a defining feature of modern agentic systems [55, 62].

Service Design Integration and Participatory Refinement
The prototype's design was grounded in earlier participatory sessions, persona development, and feedback gathered through UI concept testing. By visualizing the backend logic in LangGraph UI [97], and translating abstract workflows into tangible interface representations using Lovable.dev [82], the project made the system logic accessible to stakeholders and enabled discussion around its role in actual work. This co-creative approach ensured that the prototype aligned with real constraints and could be realistically integrated without disrupting existing routines. The solution thus acted not only as a technical artifact, but also as a service touchpoint, situated within a broader context of internal knowledge work.

4.3.6 Stakeholder Feedback and Informal Validation

Following the completion of the advanced prototype, a structured video walkthrough and live demonstration were presented to a group of senior stakeholders with decision-making authority within the PM function. This session also included a presentation of the user persona and a business case highlighting the strategic potential of AI-augmented research workflows. The purpose was to validate the feasibility, relevance, and value of the proposed solution and to assess its alignment with ongoing capability development initiatives.

The reception was overwhelmingly positive. Stakeholders expressed strong interest in the concept and identified several adjacent workflows where similar capabilities could be applied. As a direct outcome of this engagement, the organization has committed to continuing the development of the concept, initiating a deeper requirements enquiry phase. This next step will focus on refining technical specifications, defining integration pathways, applying modular development practices, and coordinating development efforts across teams with the intention of maturing the prototype into a deployable solution.

While this thesis did not aim to deliver a production-ready system, the outcomes suggest that the work functioned successfully as both a proof-of-concept and a strategic catalyst. This validation phase exemplifies the service design principle of co-creating value through stakeholder engagement, serving not only as a means of design iteration but also as a mechanism for building organizational momentum and buy-in for long-term innovation.

5 Discussion

The discussion section aims to critically interpret and contextualize the empirical findings of this thesis within the broader landscape of academic research and organizational practice. Throughout this discussion, *generative AI* refers to the broad class of models capable of producing novel content, such as text, code, or images. *Agentic AI* denotes a subset of generative AI distinguished by autonomous, goal-directed behavior, planning, and adaptive workflow orchestration, capabilities enabled by recent advances in LLMs and informed by both classic and contemporary theories of agentic systems. This distinction reflects the ongoing evolution of the field and is central to the findings and implications presented in this thesis.

Building on the results of a case study in a large enterprise context using a service design perspective, this section reflects on both the strengths and limitations of the study's approach, synthesizes theoretical implications, and examines practical, ethical, and strategic considerations. The purpose is to move beyond mere description, offering a nuanced account of how generative AI can augment knowledge work, and what barriers remain to realizing this potential in real-world settings.

By systematically relating the study's outcomes to extant literature on digital transformation, human-AI collaboration, and service design, the discussion seeks to identify where this research confirms, extends, or challenges prevailing theories. Particular attention is given to the organizational dynamics uncovered through empirical engagement, such as fragmented workflows, hidden pain points, and the situated nature of knowledge work, which are often overlooked in technology-driven narratives. In doing so, the section provides both a critical reflection on methodological choices and a platform for considering broader implications, including ethical responsibilities and future directions for research and practice.

5.1 Interpretation of Main Findings

A core finding of this study is the identification of manual and inefficient market and competitor research as a persistent bottleneck within knowledge-intensive product management workflows. This aligns with and extends prior work highlighting cognitive overload and fragmentation as endemic challenges in contemporary knowledge work

[13, 14]. However, the present research advances the conversation by unpacking the organizational and sociotechnical dynamics that underlie these issues. Specifically, the analysis revealed that pain points are not simply a function of inadequate tooling or insufficient automation, but are deeply intertwined with cross-functional collaboration patterns, unstructured data sources, and ad hoc knowledge-sharing routines.

The development and iterative evaluation of an advanced prototype, grounded in service design principles, demonstrated the feasibility and value of moving beyond traditional prompt-based automation. The prototype, designed in close collaboration with end-users, surfaced critical workflow requirements, such as the need for transparent research traceability, support for iterative refinement, and preservation of human decision authority. Informal validation sessions and stakeholder feedback confirmed that a well-integrated, agentic AI solution could act as a catalyst for capability-building, rather than as a replacement for professional expertise. This finding resonates with the emergent shift in the literature from "human-in-the-loop" to "human-in-control" models of AI deployment [15, 32].

Importantly, while the efficiency gains and reduction in manual workload were expected, the study uncovered more subtle forms of value and resistance. For example, the participatory co-design process revealed latent user needs related to trust, explainability, and the perceived risk of deskilling. Interviewees emphasized the importance of maintaining ownership over analytic judgments, and voiced concerns about over-reliance on opaque AI recommendations. This aligns with recent scholarship cautioning against technological determinism and calling for more context-sensitive approaches to AI adoption [26].

At the same time, the empirical findings diverged in some respects from the more optimistic claims in the AI transformation literature. While agentic AI hold significant promise, their impact is contingent on organizational readiness, quality of data, and the maturity of internal collaboration practices. The risk of merely shifting cognitive burdens, rather than removing them, was evident in workflows where new digital tools introduced their own complexity or required additional coordination.

The study both corroborates and complicates current understandings of AI's role in knowledge work. It supports the potential for agentic AI to address entrenched inefficiencies in the field of generative AI, but also highlights the need for user-centered, iterative, and contextually sensitive implementation strategies. The findings suggest that realizing

the full value of generative AI in complex enterprise environments requires ongoing attention to both technical and organizational dimensions, and a willingness to navigate the trade-offs and uncertainties inherent in digital transformation.

5.2 Theoretical Integration and Extension

The empirical results of this study offer several contributions to ongoing theoretical debates at the intersection of service design, human-AI collaboration, and the emerging paradigm of agentic AI. By applying a service design methodology within the context of generative AI adoption, the research foregrounds the importance of uncovering tacit and situated user needs, which are often neglected in technology-first implementations. This approach demonstrates the value of service design as both a practical methodology and a theoretical lens for organizational innovation in knowledge work.

Consistent with the work of Stickdorn et al. [28] and Kimbell [67], the study finds that service design practices, such as co-creation and iterative prototyping, are highly effective in surfacing latent pain points and reframing problems around user value rather than technological novelty. By embedding design activities within the lived experience of product managers, the research reveals how formal responsibilities and informal practices intersect, shaping both the opportunities and constraints for AI intervention. This contributes to an expanded theoretical understanding of how service design can support the alignment of new technologies with real-world workflows, helping to bridge the persistent gap between digital tool adoption and organizational value creation.

A second theoretical contribution arises from the study's challenge to prevailing "human-in-the-loop" models of AI deployment. While such models have dominated the literature, they often frame human participation as a minimal checkpoint or fallback mechanism. Drawing on Shneiderman [15] and Natarajan et al. [32], this research advocates a "human-in-control" orientation, in which sustained human oversight and agency are seen as foundational to responsible AI integration. The co-design and validation activities in this project provided empirical support for this framing: participants repeatedly emphasized the need for transparency, reversibility, and clear boundaries of automation. These insights reinforce the argument that meaningful human involvement must extend beyond mere supervision, encompassing the capacity to interrogate, override, and contextualize AI outputs within broader decision processes.

The findings also substantiate recent claims in the literature that agentic AI represents a distinctive and potentially transformative evolution of generative AI in organizational contexts [22, 23]. Unlike conventional prompt-based tools, agentic AI integrate planning, tool orchestration, and memory, enabling more adaptive and context-aware automation. In this study, the agentic AI prototype demonstrated not only efficiency improvements but also the capacity to scaffold team cognition, maintain traceability, and support knowledge reuse. These properties are in line with socio-technical systems (STS) theory, which emphasizes the mutual shaping of technology and social practices [99]. By embedding agentic AI into daily routines, the research observed early signs of distributed cognition, where the boundaries between human expertise and algorithmic support become more fluid and collaborative.

The integration of service design, human-centered AI, and agentic system perspectives in this study opens new avenues for theoretical synthesis. For instance, the findings resonate with work practice studies [29] and the design of transactive memory systems [66], highlighting how expertise is constructed, shared, and maintained in hybrid human-AI settings. Such perspectives underscore the need for continuous methodological and organizational adaptation as agentic AI becomes increasingly embedded in knowledge work.

5.3 Methodological Reflections and Limitations

This study was designed to balance methodological rigor with practical relevance, applying a qualitative single-case approach rooted in service design and the ISO 9241-210 human-centered design standard [73, 28]. Several strengths and limitations emerged in the process, which have implications for both the credibility of the findings and their generalizability.

A notable strength of the research lies in its systematic and transparent approach to qualitative data analysis. While service design frequently relies on rapid, heuristic, or so-called “quick and dirty” methods to surface user needs and prototype solutions [28], this project deliberately adopted a more structured and traceable approach. Specifically, the use of open, axial, and selective coding, combined with a transparent, two-level categorization framework based on the organization´s adoption of SAFe, was intended to ensure both depth and credibility in a complex enterprise context [90, 74]. Involving

participants throughout the process, via semi-structured interviews, artifact sharing, and co-design activities, also contributed to the depth and validity of the insights [27, 71].

This departure from conventional practice was motivated by several considerations. First, the organizational environment for this study was characterized by a high degree of complexity, legacy processes, and stakeholder scrutiny. In this setting, it was essential to generate findings that would be perceived as both actionable and trustworthy by a diverse group of decision-makers. Second, the integration of AI systems into core knowledge work presents significant risks around unintended consequences, bias, and user acceptance. A more rigorous analytical process was needed to trace conclusions back to empirical data, and to facilitate transparent dialogue with organizational stakeholders about how user needs were identified and prioritized. Third, this approach was intended to bridge the expectations of both academic and practitioner audiences, supporting generalization and transferability beyond the immediate case.

Reflecting on this methodological choice, the structured analysis enabled a deeper, more systematic mapping of pain points and opportunities than would have been possible through rapid synthesis alone. It strengthened the credibility and traceability of the research and supported effective communication with both technical and non-technical audiences. However, this rigor came at the cost of increased resource demands and a slower iteration cycle, potentially limiting the speed of prototyping and the number of design alternatives that could be explored. In practice, it required balancing analytical depth with the participatory, iterative ethos of service design, and remaining vigilant about not letting process outweigh user engagement.

The integration of AI tools, specifically Microsoft Copilot, for structuring interview transcripts represented a methodological innovation that improved analytical efficiency. Automated extraction and categorization of activity-level data reduced manual overhead and enabled rapid iteration. However, this approach also introduced potential risks related to model bias, hallucination, and overfitting to patterns not actually present in the source material. To mitigate these risks, all AI-generated outputs were manually validated and revised, and only grounded, verifiable insights were included in the analysis [50]. Nevertheless, the experience highlights the need for robust human oversight and critical review when integrating AI-assisted methods into qualitative research workflows.

Despite these strengths, the study is subject to several important limitations. The use of a single-case design, focused on one business unit within a single enterprise, con-

strains the generalizability of the findings [72]. While the purposive sampling strategy ensured relevance to the intended user group, it may have introduced bias and limited the diversity of perspectives, particularly regarding cross-functional collaboration or alternative organizational cultures. The reliance on a single, male-gendered persona for the prototyping phase is another limitation. Although this approach provided a clear and relatable focal point for design, it insufficiently captured the diversity and complexity of product management roles. Future research would benefit from the inclusion of multiple personas reflecting a broader range of identities and collaboration patterns [69, 71].

Finally, while the participatory and iterative prototyping approach proved effective in surfacing tacit knowledge and generating stakeholder buy-in, sustaining engagement over time was challenging, especially in a large, distributed enterprise environment. Competing priorities, resource constraints, and geographic location of participants all affected the continuity and depth of participation. These realities point to the intrinsic difficulties of scaling co-design and participatory innovation processes within operationally complex organizations. Furthermore, the study did not evaluate the proof of concept (PoC) over a longer time horizon. While short-term feedback indicated positive engagement and perceived usefulness, it remains unclear whether these effects would persist once the novelty factor subsides. Longitudinal evaluation would be necessary to examine whether the PoC continues to deliver value, maintains a positive user experience, and integrates sustainably into everyday workflows. This limitation points to the importance of assessing not only the immediate impact of new AI-enabled systems, but also their long-term usability, adoption, and social sustainability within enterprise contexts.

Taken together, these methodological reflections highlight both the promise and the trade-offs of combining rigorous, user-centered research methods with emerging AI-assisted analytical techniques in organizational studies. The experience suggests practical recommendations for future work, including greater investment in diversity and inclusivity during persona development, ongoing validation of AI-supported analysis, and the design of scalable participatory processes tailored to enterprise environments. Ultimately, the methodological rigor and participatory approach taken in this study support both the credibility of the findings and their practical relevance for organizations seeking to translate AI potential into operational value.

5.4 Practical Implications and Value Creation

The findings of this study have direct implications for organizations seeking to adopt generative, and in particular agentic AI, within knowledge-intensive domains. Rather than delivering a fully formed technical solution, the project functioned as a catalyst for organizational learning and capacity-building. By engaging stakeholders throughout the design and prototyping process, the research fostered early buy-in, surfaced hidden requirements, and built a shared understanding of both the possibilities and limitations of generative AI in practice.

A key practical insight concerns the role of pilot projects and early-stage prototypes. The agentic AI prototype developed during this study was not positioned as a definitive solution, but as a proof of concept intended to anchor internal discussions, clarify requirements, and identify organizational enablers and barriers. This "catalyst" role proved essential: it allowed the organization to experiment with new capabilities in a controlled setting, to reflect on workflows and data structures, and to evaluate the fit between emerging AI technologies and established ways of working. Feedback from stakeholders indicated that the prototype provided a tangible starting point for further exploration, helping to bridge the gap between abstract technological potential and concrete business needs.

The study also highlights several strategic considerations for organizations weighing the choice between developing custom AI solutions and adopting external vendor offerings. Findings suggest that while external solutions can offer speed and maturity, particularly for standardized or commoditized tasks, internal development may be justified when workflows are highly specialized, data is sensitive, or organizational differentiation depends on unique knowledge processes. A mixed approach may be optimal, with organizations leveraging vendor tools where appropriate, but investing in tailored solutions for strategically critical or under-served needs. The evaluation of "buy versus build" should thus be grounded in a clear understanding of core competencies, integration requirements, and the broader technology landscape [10, 3].

Successful and scalable AI adoption was found to rely not only on technical feasibility, but also on a set of organizational prerequisites: trust, transparency, skill development, robust data governance, and thoughtful change management. Building trust requires involving end-users in the design and validation of AI tools, ensuring that systems operate with

clear boundaries and provide traceable outputs. Transparency, in both algorithmic logic and decision processes, is vital for acceptance, as is investment in upskilling employees to collaborate effectively with AI systems [15, 26]. Data quality and governance emerged as foundational issues: without clean, accessible, and well-structured data, even the most advanced AI systems are likely to under-perform or introduce new risks. Change management practices, including early stakeholder engagement, clear communication of benefits and limitations, and iterative roll-out strategies, are essential for building momentum and sustaining adoption over time [8].

Taken together, the project demonstrates that the organizational value of agentic AI lies not only in automating repetitive tasks, but in serving as a focal point for strategic learning and capability-building. The process of co-design, prototyping, and reflective evaluation prepares the ground for more ambitious AI initiatives and supports the development of a more adaptive, user-centered innovation culture.

5.5 Ethical Reflections

The implementation of generative AI, including agentic AI, in knowledge-intensive work environments raises a number of ethical considerations that extend beyond technical performance or process efficiency. Central among these are questions of privacy, data protection and bias. As organizations increasingly integrate AI systems into sensitive workflows, the risk of unintended consequences, including the exposure of confidential information, the amplification of existing biases, and the automation of poor practices, becomes more pronounced [56, 50].

This study adopted a “human-in-control” orientation throughout the design and prototyping phases, which proved critical in addressing many of these concerns. By embedding human oversight into each stage of the research assistant’s operation, the system maintained a clear boundary between automation and judgment. Users were not only able to review, modify, or reject AI-generated outputs, but were also given transparency into the sources and logic underlying those outputs. This approach aligns with emerging recommendations for responsible AI, which emphasize the necessity of maintaining meaningful human agency and ensuring that sensitive decisions remain within the purview of human professionals [15, 32].

Despite these safeguards, the study surfaced unresolved dilemmas around the appro-

priate boundaries of AI use in knowledge work. Certain tasks, such as those involving ethical judgment, highly contextual decision-making, or creativity, proved resistant to automation, reinforcing the view that not all aspects of knowledge work are suitable for AI augmentation. There is an ongoing need to carefully delineate the scope of AI interventions, ensuring that automation supports rather than supplants human expertise, especially where individual rights, fairness, or organizational accountability are at stake.

Finally, the risks associated with bias and data quality remain significant. Even well-intentioned AI systems may inadvertently reproduce or amplify existing inequities if not carefully monitored and audited. This challenge is compounded by the opacity of many large language models, which can make it difficult to trace or explain specific outputs [53]. As a result, ethical AI adoption must be supported by robust governance frameworks, regular impact assessments, and a culture of critical scrutiny within organizations. In conclusion, responsible deployment of AI systems in knowledge work requires sustained attention to privacy, transparency, and fairness, underpinned by mechanisms that ensure ongoing human oversight and the capacity to intervene when necessary.

5.6 Critical Perspective on the AI Hype

The rapid proliferation of agentic AI has been accompanied by bold claims regarding its potential to transform productivity, creativity, and decision-making in knowledge-intensive organizations. However, the findings of this study underscore the persistent gap between technological promise and realized value, a dynamic widely recognized as the "productivity paradox" [26, 14]. Despite substantial investments in digital tools, many organizations continue to struggle with fragmented workflows, manual coordination, and limited gains in actual effectiveness.

Several factors help explain this paradox. First, technology-driven approaches often underestimate the complexity of organizational routines, the entrenchment of legacy processes, and the tacit dimensions of expertise that resist codification. As a result, AI initiatives that fail to account for user context and organizational realities frequently stall at the proof-of-concept stage or generate only incremental improvements [13]. Second, there is a tendency toward "AI-washing," where projects are branded as AI-driven without meaningful integration into core business practices. This not only dilutes the strategic impact of AI investments but can also erode trust among stakeholders when anticipated

benefits fail to materialize [10, 8].

The present research suggests that genuine productivity gains from AI depend less on the raw capabilities of the technology, and more on context-sensitive integration, user-centered design, and iterative refinement. By anchoring the adoption process in lived workflows and participatory methods, organizations can better identify where AI offers true leverage and where it may simply reinforce existing inefficiencies. This approach stands in contrast to overly optimistic narratives that present AI as a turnkey solution, rather than a catalyst for organizational learning and transformation. Furthermore, the barriers to adoption identified in this study, ranging from data quality and governance challenges to cultural resistance and skill gaps, highlight the limitations of a purely technical perspective on AI transformation. Sustainable impact requires investment in organizational capabilities, clear communication of benefits and limitations, and the cultivation of critical reflection around both the opportunities and the risks presented by new technologies.

In short, a critical perspective on the AI hype demands recognition of both the structural and cultural factors that shape real-world outcomes. Rather than chasing inflated promises, organizations should prioritize thoughtful experimentation, transparent evaluation, and ongoing alignment of AI initiatives with evolving human and business needs.

5.7 Directions for Future Research and Practice

The pace of advancement in agentic AI during the course of this project has been extraordinary, underscoring the need for both adaptive research strategies and organizational agility. From January to June alone, the emergence of more sophisticated model coordination platforms (MCPs), advances in orchestration frameworks, and early signals of quantum-enhanced AI capabilities have redefined the technological landscape and opened new avenues for enterprise application.

Looking forward, research must address not only the practicalities of longitudinal deployment, evaluating the sustained impact of agentic AI in real operational environments, but also the implications of rapidly evolving architectures and platforms. There is a pressing need to investigate how multi-agent and multi-modal orchestration frameworks can be leveraged to enable more resilient, context-aware, and scalable AI solutions. As organizations begin to experiment with MCPs that coordinate the activities of diverse

models, new questions emerge around governance, interoperability, and emergent system behavior.

Quantum computing represents another frontier with potentially disruptive consequences for AI in knowledge work. While still at an early stage, quantum-enabled models could soon extend the capacity of agentic AI systems for large-scale optimization, reasoning, and simulation. Research is needed to anticipate the organizational, ethical, and security challenges associated with quantum-accelerated AI, and to develop new evaluation frameworks that can keep pace with the technical complexity and speed of change.

Given this dynamic context, future studies should also examine the interplay between technology adoption cycles and organizational learning. Comparative evaluations of custom-built versus vendor-provided solutions remain crucial, but must now account for shifting market standards, open-source ecosystems, and the modularization of AI capabilities. Furthermore, as regulatory landscapes and best practices for responsible AI evolve in parallel, the ability to rapidly adapt governance, transparency, and validation mechanisms will become a core differentiator for enterprise success.

Practitioners and researchers alike should embrace iterative, anticipatory approaches, remaining vigilant to emerging trends, actively engaging with interdisciplinary expertise, and prioritizing adaptability over static planning. As the boundaries of what is possible continue to shift, organizations that invest in flexible architectures, continuous upskilling, and participatory implementation strategies will be best positioned to harness the transformative potential of agentic AI.

5.8 Closing Reflection

This thesis set out to explore how generative AI could meaningfully augment knowledge work in the context of enterprise product management. While the initial scope was broad, the empirical findings and iterative prototyping process revealed that the most promising opportunities for impact emerged not from generic, prompt-based automation, but from the more advanced integration of agentic AI systems. In this way, the research journey itself reflects the ongoing shift in the field, from seeing generative AI as a standalone tool, to understanding its true potential as part of orchestrated, context-aware agentic workflows.

A key contribution of this work is the demonstration that rigorous, user-centered, and participatory methods can surface both the technical and organizational prerequisites for effective AI augmentation. By adopting a structured approach to qualitative analysis, uncommon in typical service design, the study generated findings that are both credible and actionable for practitioners navigating complex enterprise realities. The development of an advanced prototype of the agentic AI system served less as an end-point and more as a catalyst for learning, dialogue, and capability-building within the organization.

Importantly, the research highlights that the value of agentic AI lies not only in efficiency gains, but in supporting new forms of distributed cognition, knowledge reuse, and adaptive decision-making. At the same time, the findings reinforce that effective and responsible AI implementation depends on sustained human oversight, transparency, and a clear delineation of boundaries between automation and human judgment.

Ultimately, the study provides a foundation for future research and practical experimentation, showing how context-sensitive, participatory, and analytically rigorous approaches can help organizations realize the full potential of AI-driven transformation. As digital systems evolve, it will be increasingly important to keep human expertise and organizational learning at the center, ensuring that AI functions not as a replacement, but as a collaborative partner in shaping the future of knowledge work.

6 Conclusion

This study set out to investigate how generative AI can augment knowledge work within a large enterprise context, using a service design perspective. It was guided by three objectives: (1) to identify recurring workflow challenges and user needs in knowledge-intensive work that may be addressable through generative AI; (2) to evaluate and prioritize opportunity areas for generative AI intervention; and (3) to design and develop an advanced prototype that demonstrates meaningful AI augmentation within a selected workflow.

The qualitative single-case study revealed that persistent pain points are deeply embedded in everyday routines and informal practices of distributed teams, reflecting complex social and systemic factors rather than purely technical limitations. This highlights that effective augmentation requires attention to context and human factors, not just the deployment of new tools. A central finding was that the greatest potential for generative AI lies not in automating isolated tasks, but in orchestrating adaptive, agentic workflows that complement and extend human judgment.

The development and validation of an agentic AI research assistant demonstrated how generative AI, when embedded within agentic system architectures, can move beyond prompt-based applications. Such systems enable more transparent, context-aware, and goal-directed support for knowledge workers, while also maintaining the necessary boundaries of human oversight and control. Feedback from organizational stakeholders confirmed the value and feasibility of this approach. Importantly, the prototype served not as a finished solution, but as a catalyst for organizational learning and reflection. It enabled participants to reimagine more effective ways of working, surfacing both new opportunities for AI augmentation and practical constraints, such as the ongoing need for explainability, trust, and fit with established practices. This process reinforced the need for participatory, iterative implementation strategies and highlighted that lasting impact depends on aligning technical interventions with organizational values and routines.

Theoretically, this research advances ongoing debates about the role of generative and agentic AI in knowledge-intensive environments. It demonstrates that user-centered approaches, incorporating service design methodologies, are essential for surfacing tacit needs, fostering organizational buy-in, and bridging the gap between digital transformation rhetoric and real value creation. The thesis further distinguishes between generative

AI as an enabling technology and agentic systems as the practical mechanism for orchestrating complex workflows, clarifying the conditions under which each delivers value.

Methodologically, the work contributes by integrating rigorous qualitative analysis grounded in open, axial, and selective coding, with participatory co-design and multi-criteria opportunity evaluation. This combination ensured analytical depth, transparency, and practical relevance, supporting both the credibility of the findings and their transferability to similar contexts.

For practitioners, the results suggest that successful AI augmentation depends on more than technical capability. Participatory and iterative implementation, robust data governance, continuous upskilling, and sustained attention to the social dimensions of change are all critical enablers. Enterprises should approach AI adoption not simply as the introduction of new tools, but as an evolving, collaborative process rooted in real human needs and adaptive to ongoing change. The use of structured opportunity evaluation frameworks can further support organizations in prioritizing investments where AI can deliver the greatest leverage, balancing impact, feasibility, risk, and scalability.

Like any study, this research has limitations. The single-case design and qualitative approach constrain generalizability, the technology landscape for agentic AI continues to evolve rapidly and practical challenges remain in scaling participatory design and ensuring diversity of perspectives. Nevertheless, the findings offer a robust foundation for further research and experimentation, both within the organization studied and for others seeking to navigate the complexities of advanced AI-driven knowledge work augmentation.

To conclude, this thesis demonstrates that meaningful augmentation of knowledge work with generative AI is best achieved through the thoughtful development of agentic systems, the application of participatory and user-centered methods grounded in service design principles, and sustained organizational learning. Rather than treating AI as a one-off technical fix, organizations should approach its adoption as a journey, one that is collaborative, iterative, and deeply attuned to the evolving landscape of technology, design, and knowledge work. These insights lay a pathway for more actionable, nuanced, and human-centered AI integration in complex enterprise settings, and set a clear direction for future research and practice.

References

[1] Stanford Institute for Human-Centered Artificial Intelligence. *Artificial Intelligence Index Report 2025*. Accessed: 2025-06-05. 2025. URL: https://hai.stanford.edu/ai-index.

[2] *Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations*. Accessed: 2025-06-05. Gartner. May 2024. URL: https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations.

[3] McKinsey Global Institute. *The Economic Potential of Generative AI: The Next Productivity Frontier*. Accessed: 2025-06-05. 2024. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

[4] OpenAI. *Introducing GPT-4o (and GPT-4.5): OpenAI's Next-Generation Model*. Accessed: 2025-06-05. May 2024. URL: https://openai.com/index/introducing-gpt-4-5/?utm_source=chatgpt.com.

[5] Cade Metz and Mitchell Clark. *OpenAI Strikes Deal With Jony Ive for New AI Hardware*. Accessed: 2025-06-05. May 2025. URL: https://www.nytimes.com/2025/05/21/technology/openai-jony-ive-deal.html.

[6] *Europe breaks another record for VC investment in GenAI*. Accessed: 2025-06-05. PitchBook. May 2025. URL: https://pitchbook.com/news/articles/europe-breaks-another-record-for-vc-investment-in-gen-ai.

[7] *Worldwide Artificial Intelligence Spending Guide*. Accessed: 2025-06-05. International Data Corporation (IDC). 2024. URL: https://my.idc.com/getdoc.jsp?containerId=IDC_P33198.

[8] *State of Generative AI in the Enterprise: Now Deciding, Experimenting, and Implementing at Scale*. Accessed: 2025-06-05. Deloitte. 2024. URL: https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html.

[9] Erik Brynjolfsson, Daniel Rock, and Chad Syverson. *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*. Accessed: 2025-06-05. 2018. URL: https://ide.mit.edu/sites/default/files/publications/IDE%20Research%20Brief_v0118.pdf.

[10] *Where's the Value in AI?* Accessed: 2025-06-05. Boston Consulting Group (BCG). 2024. URL: https://www.bcg.com/publications/2024/wheres-value-in-ai.

[11] *AI Linked to a Fourfold Increase in Productivity Growth*. Accessed: 2025-06-05. PwC. 2025. URL: https://www.pwc.com/gx/en/news-room/press-releases/2025/ai-linked-to-a-fourfold-increase-in-productivity-growth.html.

[12] Melissa Webster and George Westerman. "Generate Value From GenAI With'Small t'Transformations". In: *MIT Sloan Management Review (Online)* (2025), pp. 15–21.

[13] Allison Woodruff et al. "How knowledge workers think Generative AI will (not) transform their industries". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–26.

[14] Fabrizio Dell'Acqua et al. "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality". In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013 (2023).

[15] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[16] Konrad Sowa, Aleksandra Przegalinska, and Leon Ciechanowski. "Cobots in knowledge work: Human–AI collaboration in managerial professions". In: *Journal of Business Research* 125 (2021), pp. 135–142.

[17] Stan Franklin and Art Graesser. "Is it an agent, or just a program?: A taxonomy for autonomous agents". In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer. 1996, pp. 21–35.

[18] Pat Langley, John E Laird, and Seth Rogers. "Cognitive architectures: Research issues and challenges". In: *Cognitive Systems Research* 10.2 (2009), pp. 141–160.

[19] Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994.

[20] John R Anderson et al. "An integrated theory of the mind." In: *Psychological Review* 111.4 (2004), p. 1036.

[21] Rodney A Brooks. “Intelligence without representation”. In: *Artificial Intelligence* 47.1-3 (1991), pp. 139–159.

[22] Aditi Singh et al. “Enhancing AI systems with agentic workflows patterns in large language model”. In: *2024 IEEE World AI IoT Congress (AIIoT)*. IEEE. 2024, pp. 527–532.

[23] Yuheng Cheng et al. “Exploring large language model based intelligent agents: Definitions, methods, and prospects”. In: *arXiv preprint arXiv:2401.03428* (2024).

[24] Peter F Drucker. “Knowledge-worker productivity: The biggest challenge”. In: *California Management Review* 41.2 (1999), pp. 79–94.

[25] Thomas H Davenport. *Thinking for a living: How to get better performances and results from knowledge workers*. Harvard Business Press, 2005.

[26] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. “Generative AI at work”. In: *The Quarterly Journal of Economics* (2025), qjae044.

[27] Jan Gulliksen et al. “Key principles for user-centred systems design”. In: *Behaviour and Information Technology* 22.6 (2003), pp. 397–409.

[28] M Hormess et al. “This Is Service Design Doing: Applying Service Design Thinking in the Real World”. In: *A Practitioner’s Handbook* (2018).

[29] Wanda J Orlikowski. “Knowing in practice: Enacting a collective capability in distributed organizing”. In: *Organization Science* 13.3 (2002), pp. 249–273.

[30] Kelly B Wagman, Matthew T Dearing, and Marshini Chetty. “Generative AI Uses and Risks for Knowledge Workers in a Science Organization”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025, pp. 1–17.

[31] Yogesh K Dwivedi et al. “Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy”. In: *International Journal of Information Management* 57 (2021), p. 101994.

[32] Sriraam Natarajan et al. “Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 27. 2025, pp. 28594–28600.

[33] Arthur L Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229.

[34] Batta Mahesh et al. “Machine learning algorithms-a review”. In: *International Journal of Science and Research (IJSR).[Internet]* 9.1 (2020), pp. 381–386.

[35] Wayne Xin Zhao et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* 1.2 (2023).

[36] Humza Naveed et al. “A comprehensive overview of large language models”. In: *ACM Transactions on Intelligent Systems and Technology* 16.5 (2025), pp. 1–72.

[37] Brenden M Lake et al. “Building machines that learn and think like people”. In: *Behavioral and Brain Sciences* 40 (2017), e253.

[38] Jingfeng Yang et al. “Harnessing the power of LLMs in practice: A survey on Chatgpt and beyond”. In: *ACM Transactions on Knowledge Discovery from Data* 18.6 (2024), pp. 1–32.

[39] Ashish Vaswani et al. “Attention is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.

[40] Mohaimenul Azam Khan Raiaan et al. “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges”. In: *IEEE Access* 12 (2024), pp. 26839–26874.

[41] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[42] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[43] Yunfan Gao et al. “Retrieval-augmented generation for large language models: A survey”. In: *arXiv preprint arXiv:2312.10997* 2 (2023).

[44] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[45] Lei Wang, Wanyu Xu, et al. “Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models”. In: *arXiv preprint arXiv:2305.04091* (2023).

[46] Benfeng Xu et al. “Expertprompting: Instructing large language models to be distinguished experts”. In: *arXiv preprint arXiv:2305.14688* (2023).

[47] Shunyu Yao et al. “Tree of thoughts: Deliberate problem solving with large language models”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 11809–11822.

[48] Maciej Besta et al. “Graph of thoughts: Solving elaborate problems with large language models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17682–17690.

[49] Yue Zhang et al. “Siren’s song in the AI ocean: a survey on hallucination in large language models”. In: *arXiv preprint arXiv:2309.01219* (2023).

[50] Ziwei Ji et al. “Survey of hallucination in natural language generation”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[51] Jean Kaddour et al. “Challenges and applications of large language models”. In: *arXiv preprint arXiv:2307.10169* (2023).

[52] Lichao Sun et al. “Trustllm: Trustworthiness in large language models”. In: *arXiv preprint arXiv:2401.05561* 3 (2024).

[53] Amina Adadi and Mohammed Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160.

[54] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[55] Joon Sung Park et al. “Generative agents: Interactive simulacra of human behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023, pp. 1–22.

[56] Junaid Raiaan and et al. “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges”. In: *arXiv preprint arXiv: 2402.05438* (2024).

[57] Zhengbao Jiang et al. “How can we know what language models know?” In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 423–438.

[58] Yujia Chang et al. “A survey on evaluation of large language models”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 15.3 (2024), pp. 1–45.

[59] Vincent C Müller and Nick Bostrom. *Fundamental issues of artificial intelligence*. Vol. 376. Springer, 2016.

[60] William J. Clancey. *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge: Cambridge University Press, 1993.

[61] Klaus Mainzer. "Toward a theory of intelligent complex systems: from symbolic AI to embodied and evolutionary AI". In: *Fundamental Issues of Artificial Intelligence* (2016), pp. 241–259.

[62] Yongliang Shen et al. "Hugginggpt: Solving AI tasks with Chatgpt and its friends in Hugging Face". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 38154–38180.

[63] Yaobo Liang et al. "Taskmatrix. AI: Completing tasks by connecting foundation models with millions of APIs". In: *Intelligent Computing* 3 (2024), p. 0063.

[64] Shishir G Patil et al. "Gorilla: Large language model connected with massive APIs". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 126544–126565.

[65] Lei Wang et al. "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6 (2024), p. 186345.

[66] Konstantin Hopf et al. "The group mind of hybrid teams with humans and intelligent agents in knowledge-intense work". In: *Journal of Information Technology* 40.1 (2025), pp. 9–34.

[67] Lucy Kimbell. "Rethinking design thinking: Part I". In: *Design and Culture* 3.3 (2011), pp. 285–306.

[68] Stefan Holmlid. "Interaction design and service design: Expanding a comparison of design disciplines". In: *Nordic Design Research* 2 (2007).

[69] Elizabeth B-N Sanders and Pieter Jan Stappers. "Co-creation and the new landscapes of design". In: *Co-Design* 4.1 (2008), pp. 5–18.

[70] Marc Steen. "Benefits of Co-design in Service Design Projects". In: *International Journal of Design* 5.2 (2013). Reaffirmed in expanded findings, pp. 53–60.

[71] Marc Steen, Maarten Manschot, and Nicole De Koning. "The role of co-design in service design projects". In: *International Journal of Design* 5.2 (2011), pp. 53–60.

[72] Robert K Yin. *Case study research and applications: Design and methods*. Sage publications, 2017.

[73] International Organization for Standardization. *ISO 9241-210:2019 Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems*. Accessed: 2025-05-14. 2019. URL: https://www.iso.org/standard/77520.html.

[74] Scaled Agile Inc. *SAFe Framework*. https://framework.scaledagile.com/. Accessed: 2025-05-24. 2025.

[75] Dean Leffingwell. *SAFe® 4.0 reference guide: scaled agile framework® for lean software and systems engineering*. Addison-Wesley Professional, 2016.

[76] Scaled Agile Inc. *Product Management - SAFe Framework*. https://framework.scaledagile.com/product-management/. Accessed: 2025-05-24. 2025.

[77] Hugh Beyer and Karen Holtzblatt. "Contextual inquiry". In: *Defining Customer-Centered Systems* 31 (1998).

[78] Barbara DiCicco-Bloom and Benjamin F Crabtree. "The qualitative research interview". In: *Medical Education* 40.4 (2006), pp. 314–321.

[79] William C Adams. "Conducting semi-structured interviews". In: *Handbook of Practical Program Evaluation* (2015), pp. 492–505.

[80] Valerie Belton and Theodor Stewart. *Multiple criteria decision analysis: an integrated approach*. Springer Science & Business Media, 2012.

[81] José Figueira, Salvatore Greco, and Matthias Ehrgott. *Multiple criteria decision analysis: State of the art surveys*. Springer Science & Business Media, 2005.

[82] Lovable Inc. *Lovable.dev — Build software products using only a chat interface*. 2025. URL: https://lovable.dev (visited on 05/20/2025).

[83] LangChain. *LangGraph: A library for building agentic, stateful LLM workflows*. Accessed: 2024-06-05. 2024. URL: https://www.langchain.com/langgraph.

[84] Michael Quinn Patton. *Qualitative research & evaluation methods*. sage, 2002.

[85] Lawrence A Palinkas et al. "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research". In: *Administration and Policy in Mental Health and Mental Health Services Research* 42 (2015), pp. 533–544.

[86] Miro. *About Miro*. Accessed: 2024-06-05. 2024. URL: https://miro.com/about/.

[87] Abbie Griffin and John R Hauser. “The voice of the customer”. In: *Marketing Science* 12.1 (1993), pp. 1–27.

[88] Svend Brinkmann. *Qualitative interviewing*. Oxford University Press, 2013.

[89] Philipp Mayring. “Qualitative content analysis: theoretical foundation, basic procedures and software solution”. In: *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (2014).

[90] Juliet Corbin and Anselm Strauss. *Basics of qualitative eesearch: Techniques and procedures for developing grounded theory*. 3rd. Thousand Oaks, CA: Sage Publications, 2008.

[91] Margrit Schreier. *Qualitative content analysis in practice*. London: Sage Publications, 2012.

[92] Kevin Mullet and Darrell Sano. *Designing Visual Interfaces: Communication oriented techniques*. Prentice Hall, 1995.

[93] OpenAI. *API Platform*. URL: `https://openai.com/api/` (visited on 05/20/2025).

[94] Mendable AI. *Firecrawl - Turn websites into LLM-ready data*. 2025. URL: `https://www.firecrawl.dev` (visited on 05/20/2025).

[95] LangChain AI. *Open Deep Research: An open-source research assistant*. `https://github.com/langchain-ai/open_deep_research`. Accessed: 2025-05-20. 2025.

[96] LangChain AI. *LangChain - The platform for reliable agents*. 2025. URL: `https://www.langchain.com` (visited on 05/20/2025).

[97] LangChain AI. *LangGraph Studio - Specialized agent IDE documentation*. 2024. URL: `https://langchain-ai.github.io/langgraph/concepts/langgraph_studio/` (visited on 05/20/2025).

[98] Vetenskapsrådet. *Good Research Practice (God forskningssed)*. Swedish Research Council. 2017. URL: `https://www.vr.se/english/analysis/reports/our-reports/2017-08-31-good-research-practice.html`.

[99] Fred E Emery and Eric L Trist. “Socio-technical systems”. In: *Management Science, Models and Techniques* 2 (1960), pp. 83–97.

Appendix A. Interview Guide

The following guide was used to conduct all semi-structured interviews with product managers and product marketing managers as part of this study. All organization-specific references have been anonymized for confidentiality.

Objectives:

- Understand how product management professionals allocate their time and identify key activities
- Identify areas where AI-driven solutions or data analytics could provide meaningful support
- Examine workflows, pain points, and opportunities for improvement in current practice

Interview Themes:

1. Responsibilities and activities within product management
2. Tools and platforms used in daily work
3. Activities participants wished they had more time or support with; challenges in obtaining customer and market insights
4. Perceived opportunities, concerns, and success criteria for AI and analytics in product management
5. If time allowed: How market and customer insights influence product strategies
6. If time allowed: Decision-making process and development workflow
7. Invitation to participate in testing of a future AI-supported prototype

Probing Areas (used when needed):

- Role background, tenure, main tasks, typical week, key stakeholders
- Backlog management and collaboration tools; experience with AI or automation
- Handling of documentation and data; causes of work delays; impact of unexpected issues
- Desired improvements, concerns about AI adoption, and success measures
- Approaches to gathering customer feedback, using personas/Voice of Customer, translating insights into requirements
- Methods for gathering requirements, communicating with development, using

metrics to assess product success

Interview Best Practices:

- Begin with open-ended questions and allow participants to guide the conversation
- Use probes as needed to clarify or explore additional detail
- Listen for unexpected workflows, tool use, pain points, and signs of enthusiasm or frustration
- Capture patterns in time allocation, decision-making, collaboration, and documentation

Appendix B. Categorization Framework

This framework outlines how pain points were categorized during analysis, based on the SAFe product management domains. Each responsibility area includes associated key activities, their descriptions, and primary source references used for alignment.

Responsibility Area 1: Other

Key Activity	Activity Description	Key Sources
Communication and administration	Administrative and cross-functional communication tasks.	
Sales & Revenue Operations	Overseeing sales processes and revenue management.	
Other	Tasks not categorized above.	

Responsibility Area 2: Market Research & Analysis

Key Activity	Activity Description	Key Sources
Conduct market research	Gather and analyze data on industry trends, customer needs, and market dynamics to inform product decisions.	SAFe Product Management, Customer Centricity
Identify target market	Define and refine customer segments based on demographics, behavior, and needs to optimize product positioning.	SAFe Product Management, Customer Centricity
Monitor market trends and competitive landscape	Continuously track industry developments and competitor activities to anticipate market shifts.	SAFe Product Management, Customer Centricity
Gather voice of customer data	Collect and synthesize feedback from customers through interviews, surveys, and analytics to identify pain points and opportunities.	SAFe Product Management, Customer Centricity
Perform opportunity identification and sizing	Evaluate potential product opportunities by assessing market size, customer demand, and business impact.	SAFe Product Management, Customer Centricity
Conduct research on regulatory and compliance standards	Review compliance-related requirements and risks.	SAFe Compliance Management

Responsibility Area 3: Product Strategy & Vision

Key Activity	Activity Description	Key Sources
Define product vision and strategy	Establish a long-term direction for the product based on market insights, customer needs, and business strategy.	SAFe Strategic Themes
Ensure alignment between product objectives and strategic themes	Translate high-level business goals into actionable product initiatives that support organizational objectives.	SAFe Strategic Themes
Develop and communicate value streams	Map how the product delivers value to customers and stakeholders across different touchpoints.	SAFe Strategic Themes
Align product vision with enterprise goals	Ensure product vision and strategy are integrated with broader company objectives and priorities.	SAFe Strategic Themes

Responsibility Area 4: Product Planning & Road Mapping

Key Activity	Activity Description	Key Sources
Create and maintain product roadmap	Develop and update a strategic plan outlining key milestones, feature releases, and dependencies.	WSJF Prioritization, SAFe PI Planning
Prioritize features using WSJF	Use economic impact analysis to rank and schedule features based on customer value and effort required.	WSJF Prioritization, SAFe PI Planning
Align backlog priorities with business objectives	Ensure the backlog reflects strategic business priorities and customer needs.	WSJF Prioritization, SAFe PI Planning
Present roadmap and feature priorities in PI planning	Communicate roadmap and feature priorities in Program Increment planning.	SAFe PI Planning
Identify dependencies across teams and ARTs	Recognize and coordinate dependencies across teams to ensure smooth execution of planned work.	SAFe PI Planning

Responsibility Area 5: Product Development & Delivery

Key Activity	Activity Description	Key Sources
Collaborate with Agile Release Trains (ARTs) for feature development	Work closely with development teams to ensure efficient feature implementation and alignment with business goals.	SAFe Agile Product Delivery, Continuous Delivery, SAFe Agile Release Trains (ARTs)
Define acceptance criteria for features	Establish clear, measurable criteria for feature requirements and value delivery.	SAFe Agile Product Delivery, Continuous Delivery, SAFe Agile Release Trains (ARTs)
Integrate development, testing, and deployment	Enable faster and more reliable product releases through integrated processes.	SAFe Agile Product Delivery, Continuous Delivery, SAFe Agile Release Trains (ARTs)
Monitor feature development and team progress	Track cycles, velocity, and progress towards release goals.	SAFe Agile Product Delivery, Continuous Delivery, SAFe Agile Release Trains (ARTs)
Bug triaging, support ticket management, and prioritization	Manage tasks and tickets according to established priorities (e.g., MoSCoW).	

Responsibility Area 6: Launch Planning & Execution

Key Activity	Activity Description	Key Sources
Define launch approach	Plan for launch of new features or products, including positioning, messaging, and customer onboarding.	SAFe PI Planning, OKR Key Results Tracking
Align cross-functional teams	Coordinate marketing, sales, support, and engineering for a successful launch.	SAFe PI Planning, OKR Key Results Tracking
Monitor adoption and feedback	Measure effectiveness and make adjustments as needed.	SAFe PI Planning, OKR Key Results Tracking
Analyze outcomes and lessons learned	Gather insights to inform future releases.	SAFe PI Planning, OKR Key Results Tracking

Responsibility Area 7: Performance Monitoring & Optimization

Key Activity	Activity Description	Key Sources
Define and track success metrics	Establish measurable indicators for product performance, adoption, and satisfaction.	Measure What Matters
Analyze user engagement data	Understand customer interactions to identify areas for improvement.	Measure What Matters
Gather and respond to customer input	Continuously refine and optimize based on user feedback.	SAFe Product Management, Continuous Learning Culture
Optimize product with data-driven insights	Use analytics, A/B testing, and feedback for iterative improvements.	Measure What Matters
Monitor cost vs. value impact of decisions	Assess financial outcomes to ensure cost-effectiveness.	SAFe Lean Portfolio Management

Responsibility Area 8: Product Evolution & Lifecycle Management

Key Activity	Activity Description	Key Sources
Manage product end-of-life decisions	Plan and execute phase-out strategies for outdated or underperforming products.	SAFe Value Streams, Feature Sunsetting Strategies
Plan feature sunset strategies	Develop deprecation plans to ensure smooth transitions for users.	SAFe Value Streams, Feature Sunsetting Strategies
Ensure backward compatibility and customer transition plans	Support legacy features and guide users through transitions.	SAFe Value Streams, Feature Sunsetting Strategies
Evaluate and evolve value streams	Continuously refine product value delivery as customer needs change.	SAFe Value Streams, Feature Sunsetting Strategies

Responsibility Area 9: Product Governance & Portfolio Management

Key Activity	Activity Description	Key Sources
Align investments with Lean Portfolio Management	Ensure funding and development align with strategic priorities and customer needs.	Lean Portfolio Management, Risk & Compliance Management
Ensure regulatory compliance and risk management	Implement and monitor compliance policies.	Lean Portfolio Management, Risk & Compliance Management
Balance demand vs. capacity across teams	Optimize workload distribution and resource allocation.	Lean Portfolio Management, Risk & Compliance Management
Manage financial forecasting for product investment	Develop and track budget forecasts to ensure growth.	Lean Portfolio Management, Risk & Compliance Management
Conduct product risk assessments and trade-off analysis	Evaluate risks and benefits to inform product decisions.	SAFe Lean Portfolio Management