CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# Enhancing the User Experience for AI-driven Complex Knowledge Systems with Natural Language Interfaces

## A UX Design Approach for Diverse Professional Users in B2B

Masters thesis in Computer science and engineering

ANNIE CLAESSON, OLIVIA FRIBERG

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

# Enhancing the User Experience for AI-driven Complex Knowledge Systems with Natural Language Interfaces

A UX Design Approach for Diverse Professional Users in B2B

ANNIE CLAESSON, OLIVIA FRIBERG

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Enhancing the User Experience for AI-driven Complex Knowledge Systems with Natural Language Interfaces: A UX Design Approach for Diverse Professional Users in B2B
A UX Design Approach for Diverse Professional Users in B2B
ANNIE CLAESSON, OLIVIA FRIBERG



Supervisor: Beata Stahre Wästberg, Department of Computer Science and Engineering
Advisor: Björn Berg Marklund, Recorded Future
Examiner: Staffan Björk, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Enhancing the User Experience for AI-driven Complex Knowledge Systems with Natural Language Interfaces: A UX Design Approach for Diverse Professional Users in B2B
A UX Design Approach for Diverse Professional Users in B2B
ANNIE CLAESSON, OLIVIA FRIBERG
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

# Abstract

While user experience (UX) in AI systems and natural language interfaces has been explored in previous research and frameworks, few studies specifically focus on the business-to-business (B2B) context. Existing guidelines and design implications have primarily been shaped by leisure focused applications where they have stemmed from business-to-consumer (B2C) contexts. In the context of designing for professional users, there is currently a need for UX design guidelines specifically targeting work related contexts.

This thesis aims to address this gap by exploring what factors should be considered when seeking to improve AI-driven complex knowledge systems with natural language interfaces. It focuses on the needs of diverse professional users in B2B contexts and examines how UX design can address these factors. The study is guided by the following research questions: (1) What factors should be taken into consideration when seeking to improve AI-driven complex knowledge systems with natural language interfaces in B2B contexts? (2) What role can UX design play in addressing factors influencing the improvement of AI-driven complex knowledge systems with natural language interfaces in B2B contexts?

This study uses a mixed-method approach where qualitative data are collected through data logs, interviews, and a survey to draw insights from users in real-world B2B scenarios working with an AI- driven system provided by a threat intelligence company.

As a result of this research, eight factors were identified to affect the user experience of natural language interfaces for information retrieval in a B2B context. Eleven design guidelines are proposed to provide UX designers guidance in designing natural language interfaces that support user control and customization, trust and transparency, and communication of the AI's abilities and limitations. The identified factors and proposed guidelines offer a foundation for future research in UX design for AI-driven systems in professional environments. This study invites for further exploration through validation across different B2B sectors and the identification of additional context specific design factors. This thesis contributes to bridge the UX research gap for AI-driven natural language interfaces in professional contexts by offering practical insights to support trust, effective information retrieval, and user friendly interactions.

Keywords: UX-design, Usability, B2B, Human computer interaction, Design principles.

# Acknowledgements

We would like to express our gratitude to the company Recorded Future for providing us this exciting thesis opportunity. For continuous support, insightful discussions, and guidance throughout the project, we would like to express a special thanks to our supervisor Björn Berg Marklund. We would also like to thank the product design team and the AI development team at Recorded Future for their support, guidance, and contribution with their expert knowledge. Lastly, we would like to acknowledge our academic supervisor Beata Stahre Wästberg for her commitment to our project. Her valuable academic guidance and encouraging support have been greatly appreciated throughout this thesis.

Annie Claesson, Olivia Friberg, Gothenburg, 2025-06-10

# Contents

# List of Abbreviations

| Abbreviation | Definition |
|---|---|
| AI | Artificial Intelligence |
| B2B | Business-to-business |
| HCXAI | Human Centered Explainable Artificial Intelligence |
| HCAI | Human Centered Artificial Intelligence |
| KRNW | Knowledge Resource Nomination Worksheet |
| NLI | Natural Language Interface |
| NLP | Natural Language Processing |
| UX | User Experience |
| XAI | Explainable Artificial Intelligence |

# List of Figures

# List of Tables

# 1

# Introduction

AI-driven complex knowledge information systems play a central role in supporting professionals within business-to-business (B2B) contexts, by facilitating knowledge discovery and informing decision-making [1]. With the rapid growth of AI-driven systems, the opportunity for effective information access and analysis has improved significantly[2], [3]. However, despite these advances, many AI-driven systems face challenges in meeting the diverse needs of professional users, particularly in terms of usability, transparency, and trust [4].

Cybersecurity is an industry where AI can have a crucial role with its computational power and capabilities to process large amounts of data, providing intelligent cybersecurity services and management [5]. This study explores how UX design can bridge the gap between system complexity and usability in B2B settings. The study will be conducted together with Recorded Future, a company that provides AI-driven solutions to detect and address potential cybersecurity threats. With solutions like automated threat detection, they aim to give organizations real-time contextual intelligence to enhance their decision-making. A central feature of the platform is a natural language interface (NLI), which enables analysts to interact with threat intelligence using conversational input. [6]

UX research has traditionally centered on consumer and leisure applications, while work related contexts have received less attention [7]. In the context of designing for professional users, there is currently a need for UX design guidelines specifically addressing work contexts and professional users [7].

## 1.1 Aim and Goal

The study focuses on collecting insights regarding the challenges and needs of diverse professional users when interacting with AI-driven complex knowledge information systems. The aim is to provide insights that align with both industry needs and academic advancements, ensuring a practical and research-driven approach to challenges within AI-driven information systems. With the gained insights the goal is to propose a set of UX guidelines adapted to support UX designers in enhancing the design and usability of such systems within a B2B context.

## 1.2 Research Questions

The study addresses the following research questions:

- **RQ1:** What factors should be taken into consideration when seeking to improve AI-driven complex knowledge systems with natural language interfaces in B2B contexts?
- **RQ2:** What role can UX design play in addressing factors influencing the improvement of AI-driven complex knowledge systems with natural language interfaces in B2B contexts?

## 1.3 Scope and Limitations

This research explores factors affecting the user experience of AI-driven knowledge systems with natural language interfaces in B2B contexts, and how they could be addressed to enhance the user experience.

To understand the complexity of a diverse user groups needs, challenges and perspectives, methods that enable a broad data collection within the thesis limited timeframe will be prioritized. Given the limited time, the study focuses on exploring how design solutions could impact user trust, system affordance and perceived effectiveness in AI-driven systems.

This thesis specifically aims to develop guidelines for NLI's designed for information retrieval, recognizing that other NLI solutions could have different user requirements and design considerations, and therefore will fall outside the scope for this thesis.

To account for the users experience of an AI system being closely intertwined with the underlying AI model and architecture, this report distinguishes between factors primarily focused on technical aspects and factors related to the interface design. The study focuses on exploring the design aspects of AI systems, therefore factors primarily related to technical aspects, such as natural language processing will fall outside the scope of this research.

A limitation for the project is the constraints regarding data privacy, which influence our study in the presentation of the result, where citations and example conversations are excluded.

## 1.4 Stakeholders

The following section details the stakeholders who have an interest in or are related to this thesis.

### 1.4.1 Chalmers University of Technology

Chalmers University of Technology has a strong interest in this thesis, as it contributes to research in user experience (UX) and usability. As a university that

values innovation and technological advancement, Chalmers benefits from studies that deepen the understanding of UX, especially in complex digital environments. The insights from this research can support future academic work and practical applications, helping to refine usability principles. Moreover, this thesis aligns with Chalmers goal of connecting research with real-world challenges, making the findings valuable for both academia and industry.

### 1.4.2 Recorded Future

This thesis is being conducted in collaboration with Recorded Future, a leading threat intelligence company that holds interest in its findings. As the research progresses, we will gain deeper insights into the needs of B2B customers within the cybersecurity domain. By building on previous research and integrating our findings, this thesis could contribute with valuable knowledge, supporting Recorded Future in the development of AI-driven solutions.

### 1.4.3 End Users

End users of the Recorded Future systems have the potential to gain value from the results of this thesis. By integrating the developed UX guidelines into the system design, the user experience could be improved.

# 2

# Background

This chapter presents the relevant background and context for the study. The chapter begins by presenting the role of UX in B2B contexts, including defining the terms UX and Usability. Following this, the conceptual background of the technical scope is presented including AI and NLI. The chapter concludes by describing Recorded Futures AI.

## 2.1 The Role of UX in B2B Contexts

Previous research in the field of user experience (UX) has primarily focused on leisure contexts [7]. According to Çalar et al. this focus have been criticized for overlooking work related contexts as the two areas vary significantly. Adaptation of design practices in leisure contexts are often applied to the development of work related tools, creating a problem area where work related tools tend to focus on hedonic experiences limiting the consideration of work environments having different motivations and expectations than leisure contexts. The research within the field UX specifically related to work context is still in an immature state, emphasizing the need for further exploration and focus. More specifically, Çalar et al. highlights the need to address the unique challenge of developing work tools in a B2B context that meets the needs of multiple stakeholders. The problem derives from finding balance between the purchaser's often more business centric requirements and the user centric needs of the end user. [7]

### 2.1.1 UX and Usability

Studies within UX research focusing on the B2B context rarely provide a clear definition of UX, an issue highlighted by Çalar [7]. Furthermore, Çalar et al. states that the lack of clear definitions of UX in work related research creates challenges for both academic research and the development of practical guidelines, contributing to the ongoing imbalance between UX research in leisure and work contexts. Since many studies in the B2B context lack clear definitions of UX, Çalar et al. also emphasize the importance of distinguishing related concepts, such as usability, from UX and clarify their meanings. [7]

#### 2.1.1.1 Definition of UX

The term UX is defined by the International Organization for Standardization (ISO) as the *"Users perceptions and responses that result from the use and/or anticipated use of a system product or service."* The perceptions and responses of the user includes *"emotions, beliefs, preferences, perceptions, comfort, behaviors, and accomplishments that occur before, during and after use."* Furthermore, user experience is the result from the combination of *"brand image, presentation, functionality, system performance, interactive behavior, and assistive capabilities of a system, product or service."* It also results from *"the internal and physical state of the user based on prior experiences, attitudes, skills, abilities and personality, as well as the context of which the system is used."* [8]

#### 2.1.1.2 Definition of Usability

The term usability is defined by the International Organization of Standardization (ISO) as to what extent a system, product or service can be used by *"specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."* The specific users, goals and context of use refers to the *"particular combination of users, goals and context of use for which usability is being considered."* [8]

## 2.2 Conceptual Background of the Technical Scope

AI-driven complex knowledge systems are advanced digital systems that embed artificial intelligence techniques, such as machine learning and natural language processing, into knowledge management to support the creation, organization, sharing, and use of information to enhance organizational outcomes. [1], [9] The following sections offer a conceptual background of AI-driven systems, Natural Language interfaces (NLI) and Recorded Futures AI.

### 2.2.1 AI

We have seen rapid growth of AI chatbots in recent years, with increasing integration into various industries and services [2], [3]. According to Meshram et al. [2], we as humans have learned to adapt to the quick and effortless way of achieving our goals with the help of AI. Due to this, our expectation of a certain standard in terms of requirements increases and becomes more difficult to meet. [2]

AI-powered chatbots, also known as conversational agents, are typically based on machine learning models that use patterns and keywords to interpret user input [2]. With the use of Natural language Processing (NLP) the chatbots manage to understand and generate text that characteristics a human, making it less demanding for the user to have a dialog with the AI as no technical knowledge about specific prompts is necessary [2], [10]. These systems can be developed to adapt and improve over time by learning from user interactions, allowing for increasingly accurate and context aware responses [2]. As the use of AI continues to expand rapidly, questions

regarding how to design user friendly, trustworthy and understandable AI interfaces emerge simultaneously [3]. Conversational agents rely on users to formulate and guide the interaction through prompts, enabling the user to steer the direction of the conversation [2]. While this gives the user the power over how the conversation might play out, it puts a certain expectation on the user's abilities to ask the right questions. This fosters new perspectives on existing challenges, where designers have to ensure a good user experience and reliable outputs for the user to find it trustworthy [3].

### 2.2.2 Natural Language Interfaces (NLI)

Natural Language Interfaces (NLI) facilitates human-computer interaction by enabling users to communicate with interfaces through text or speech using natural language [11], [12]. NLI plays a crucial role for end users as it bridges the gap between humans and computer processing which allows the user to interact with the application through text or speech in a way that does not require specific commands [10]. By merging the NLP and HCI, NLI improves the user experience to be more accessible and efficient for the user, making the digital tools more intuitive and reducing the learning curve for users [12].

A primary goal of NLI is to facilitate analytical conversations, where users need to query, filter, and manipulate data effortlessly. This possibility enables various domain experts to carry out tasks more effectively without deep technical expertise. NLI has the potential to effectively streamline documentation and reporting using report generation. Additionally, Conversational interfaces such as AI-driven assistants have proven to have a leverage NLI for supporting customers and facilitating interactive learning environments. Visualization creation is another growing area where Text-to-visualization can generate visual content from natural language descriptions. This enhances data exploration for the user while also facilitating more effective storytelling. [12]

### 2.2.3 Recorded Future AI

Recorded future [6] offers a threat intelligence platform designed to help organizations identify threats, enabling proactive actions to prevent attacks. An AI-driven intelligence graph is a central part of the platform, combining internal organizational data with external intelligence to provide real-time actionable insights. Recorded Future AI is provided as a natural language interface (NLI) through which analysts interact with the intelligence graph, making analysis and production of intelligence automated. AI conversations, AI reporting, and AI insights are the three components of the Recorded Future AI. [6]

The provided background to Recorded Futures platform is essential in this thesis, as the forthcoming analysis draws upon feedback from users of the platform. This study specifically includes two components of Recorded Future AI: AI conversations and AI reporting.

# 3

# Theory

This chapter presents a relevant theoretical background to provide a solid foundation of important aspects in the design of AI-driven complex knowledge systems and natural language interfaces (NLI). Its purpose is to outline the study within existing research by highlighting key concepts, frameworks, and challenges identified in the field. As this study aims to explore important factors and challenges, the chapter also serves as a review of previous research that highlights what has already been considered and helps guide the focus of this work.

## 3.1 Information System Success and Technology Acceptance

To understand the success and usage of information systems (IS), two theoretical frameworks can be applied: Delone and McLeans IS success model [13] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [14]. Delone and McLeans model is a widely used framework in IS research, evaluating IS success through six factors; *System quality, Information quality, Use, User satisfaction, Individual impact,* and *Organizational impact*. *System quality* refers to the system's technical performance and usability, while *information quality* focuses on the accuracy and relevance of the provided information. *Use and user satisfaction* measure how the system is utilized and perceived by users. The two final factors capture the system's impact on individual performance and the organization's overall efficiency. [13]

At the same time, UTAUT highlights the factors that influence how users accept and use technology. The framework identifies four factors that have a direct impact on user acceptance and usage behavior; *Performance expectancy, Effort expectancy, Social influence,* and *Facilitating conditions* [14].

*Performance expectancy* refers to which extent a user believes the system will improve work performance, while *effort expectancy* relates to perceived ease of use. *Social influence* describes the importance of others opinions regarding system usage, and *facilitating conditions* refer to the availability of organizational and technical support for system implementation [14].

These models provide an understanding of both IS success and user acceptance, which is central in designing AI-driven complex information systems in B2B contexts.

Delone and McLeans model can be used to assess how well the system performs and its impact on individuals and organizations, while UTAUT helps analyze the factors influencing users' willingness to adopt and integrate the system into their workflow. By combining these perspectives, design decisions can be better tailored to achieve both technical success and user acceptance in professional environments.

## 3.2 Affordance

Affordance theory provides a foundational perspective on users' interactions with technology by defining the possibilities for action that an environment or system offers. The interpretation of affordance varies, and relevant literature is spread across several disciplines. Affordance theory was originally introduced by the psychologist James Gibson describing the relationship between an actor and its environment in terms of possibilities for action [15]. This concept has been applied to technology and interaction design to understand users' perceptions and interactions with systems in order to guide the design of user interfaces and interactions [16].

In Gavers technology extension on Gibsons theory of affordance, perceptible, hidden, false, and sequential affordances are introduced. Perceptible affordances refers to those clearly signaling the intended use, while hidden affordances exist but are not instantly visible. False affordances provide false cues to the user by indicating an action opportunity that does not exist. Sequential affordances describe one affordance leading to another [16]. Norman developed the affordance concept by categorizing them as real and perceived that have had a major impact on the field of human computer interaction (HCI). Real affordances refers to the actual properties that determine how an object can be used, while perceived affordances emphasize the user's beliefs about what the object can do based on the design [15], [17].

According to Gibsons perspective, affordances exist independently of perception and are constant across various users and contexts. Norman states that affordances are designed into artifacts, and emphasize the importance of affordances being intuitive and visible to effectively lead the user behavior. Gaver recognizes that affordances exist independently while arguing that they must be created perceptible for effective interaction. [15], [16]

While Gaver and Norman particularly focus on affordances on an individual level, [15] broaden the discussion by focusing on affordances in organizational contexts. They discuss the concept of affordances from existence, perception, actualization and effect perspectives. The importance of understanding how affordances are actualized and affect organizations are emphasized in order to understand the role of IT artefacts in organizations. Affordance effects occur as the result of affordance actualization. [15]

Affordance theory provides a framework for analyzing and improving the UX in AI-driven complex knowledge information systems within B2B contexts. [16] contributes with insights on perceptible affordances in digital interfaces, while [17] provides a clear distinction between real and perceived affordances to emphasize intuitive design. Pozzi et al. [15] broadens the perspective to include organiza-

tional adoption and impact. By integrating these perspectives, UX designers can develop AI-driven information systems that are both intuitive for users and effective in supporting business knowledge processes.

## 3.3 Explainable Artificial Intelligence

Although Artificial Intelligence (AI) has existed for several years, its rapid growth and widespread adoption across various fields have increasingly complicated its comprehensibility [18]. Explainable Artificial Intelligence (XAI) is a rapidly growing field that aims to bridge the gap between the technical complexity of AI and human understanding. The goal of XAI is to create interpretable data models that benefit problem-solving tasks without neglecting humans' ability to trust the provided responses and solutions [19]. By providing users insight into the otherwise complexity of machine learning (ML) models, XAI aims to help users make more informed choices. XAI emphasizes transparency, explainability, and interpretability, which are essential for ensuring that AI systems can be understood and trusted by both experts and non-experts [18].

Transparent models and decision trees serve as key tools in XAI due to their understandable mechanisms. In addition to improving interpretability, XAI provides a theoretical foundation for promoting responsible AI. By integrating explainability with principles such as fairness and reliability, the objective is to foster broader acceptance and encourage more ethical applications of artificial intelligence [18]

### 3.3.1 Human centered explainable AI

While explainable AI has an important focus on explaining complex models for technical solutions, Human centered explainable AI (HCXAI) has a greater focus on the human factors in the interaction between the AI and the user [20]. The goal of HCXAI is to ask the question of who needs the explanation, why they need it and how they will use it in order to consider the human experience. By doing so, the HCXAI moves beyond the focus of transparency and opening up the black box to addressing the need for human centered usage of AI in everyday settings [20], [21].

Lee [21] argues that transparent models do not necessarily result in a user centered AI solution. Instead, researching intuitive human behaviors before training AI models can lead to a greater understanding of natural human behaviors. To gain a more profound understanding of how users interact and interpret explanations given by the AI, Elicitation tests can be utilized. These tests facilitate the identification of how users may employ personal inputs to describe concepts or tokens when defining their questions or prompts. After the tests have been carried out, the paper emphasizes that the new insights will benefit in customizing pre-trained language models. However, the remaining challenge lies in translating these insights into concrete implementation steps within the existing language model. Lee therefore suggests that by initializing new words and optimizing their representation, the language model can evolve into intuitive human behaviours and give the user a more personalized experience. [21]

## 3.4 Developing trustworthy systems

Various disciplines within human centered design emphasize the importance of prioritizing human centered objectives when developing AI-driven systems [22]. Several disciplines agree that ethics and trustworthiness are essential elements for AI development to ensure human trust in the system.

Liao and Sundar [3] argue that trustworthiness of an AI is not established solely by system attributes, but is communicated through *trustworthy cues*. These cues are embedded in the interface design, documentation and interactions with the user and perceived through independent judgement by each person. Therefore, trustworthiness in AI needs to be thoughtfully incorporated through the quality of the system to convey users that their trust is justified. Liao and Sundar emphasize the risk that AI systems may communicate trust signals that can create a false sense of trustworthiness. Whether it is unintentionally or by design, it can result in users believing that the systems intentions are credible when they are not. To prevent this, developers and designers must understand how users process and perceive information and account for a diverse user group where the level of technical expertise and ability to process complex information might vary. [3]

Khan et al. raises important aspects regarding transparent decision making processes as key features to inform the user about how the AI reached its answer [23], [24]. By keeping the human in the loop of the decision making process, it reduces ethical concerns for deceiving users in making decisions based on misleading information [23]. Traceability mechanisms can help accomplish this, as well as explaining the system's capabilities, limitations, and decisions in an understandable way for the stakeholder involved [24]. This connects maintaining human control and providing transparency, for enhancing trustworthiness.

One framework that shares the importance of trustworthy systems is Human centered AI (HCAI), which focuses on the synergy of automated systems and human control. HCAI is a two dimensional framework that emphasizes the need for a high level of human control within applications of high levels of automation [25, p. 47]. The goal is to develop reliable, safe, and trustworthy applications that ensure humans remain in control by enabling them to make informed decisions based on the complex information generated by automated systems. Shneiderman states that *"Machine and human autonomy are both valuable in certain contexts, but a combined strategy uses automation when it is reliable and human control when it is necessary"* [25, p. 53]. With this in mind, HCAI focuses on guiding the design choices in how to make automated systems reliable, safe and trustworthy with the human in mind. Shniederman provides the following definitions of the attributes.

*Reliable systems* aim to deliver expert responses and expected outcomes under defined conditions. The reliability is commonly ensured through verification and validation, tracking and tracing failures, and fairness and predictability. [25, p. 53]

*Safe systems:* Safety in automated systems is managed through proactive risk management and following industry standards. This involves commitment to safety, extensive failure reporting, and structured validation processes to refine operations

and prevent risks.[25, p. 54]

*Trustworthy systems:* A trustworthy system goes beyond trust as it has proven itself to be deserving of the users trust. However, it is commonly difficult for the user to determine since users often lack the skillset of assessing the trustworthiness of a complex system. [25, p. 54]

These attributes are frequently referenced in discussions surrounding AI, yet Shneiderman states that they remain challenging to measure and evaluate. Consequently, the HCAI framework aims to guide designers and researchers in developing systems that integrate these attributes while also accounting for broader design considerations [25, p. 55]. From the viewpoint of this study, this perspective can be valuable as it encourages critical reflection and facilitates the formulation of questions that may uncover additional aspects relevant to the assessment of complex AI-driven systems.

### 3.4.1 HCAI Framework

The HCAI framework focuses on redefining autonomy in AI driven systems by emphasizing human control and enhancing automation [25, p. 57]. Upon HCAI, AI and automation were often viewed from a single axis perspective where the scale went from high level of human control and low level of automation to low level of human control and high level of automation[25, p. 49]. HCAI enables a two dimensional framework where applications with high levels of automation also can have a high level of human control and is shown in Fig. 3.1.

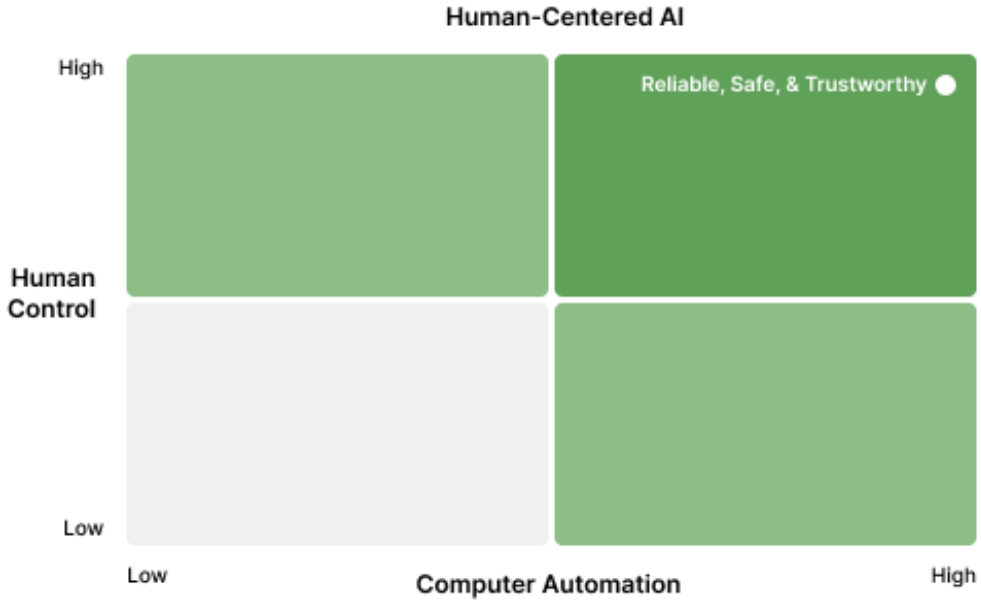


Figure 3.1: HCAI two- dimensional framework after Shneiderman.[25, p. 60]

With this perspective, the framework encourages designers and researchers to be innovative, explore new questions and reconsider existing methods to deliver AI applications that are reliable, safe and trustworthy. The framework presents two axes where human control is ranked from high to low on the left vertical axis and computer automation is ranked on the bottom horizontal axis. For systems with high automation, the ideal position is in the upper right quadrant. Here, the user has the control to override the decisions made by the AI when considered necessary. [25, p. 60]

## 3.5 Effectiveness and Efficiency in AI-driven systems

AI driven chat solutions have an impact on many industries as their capability to provide real time support enhances effectiveness and user engagement [26]. However a visible need for continuous evaluation of AI tools like chatbots is needed to make sure that the user experience and usability aspects are taken into consideration.

Effectiveness and efficiency are fundamental aspects of usability that influence how well a system helps users accomplish their goals. The ISO 9241-11 is the latest edition of an ISO standard addressing effectiveness and efficiency within the field of UX and usability [27]. Alabbas and Alomar present a framework that builds upon the ISO 9241-11 [26]. The framework defines effectiveness as the systems ability to deliver accurate and satisfactory results and is often evaluated by task success rates and error minimization. Efficiency concerns the system's capacity to support users in completing tasks with minimal effort and is often evaluated through response times and task completion rates. Together, effectiveness and efficiency are crucial in shaping the overall user experience (UX).

Systems that respond quickly while maintaining accuracy contribute to a seamless user experience, reducing frustration and cognitive strain. AI-driven systems, such as those utilizing machine learning and natural language processing, must carefully balance speed and precision to optimize user outcomes. If an AI system responds instantly but frequently makes errors, user trust decreases.

By advising professionals to measure completion time, error rates, cognitive effort, user trust and satisfaction, the framework seeks to help extract valuable information on how to build AI solutions that work seamlessly and engage the users. [26]

### 3.5.1 Measuring efficiency and effectiveness

As efficiency and effectiveness have previously been evaluated using quantitative performance metrics in AI-driven systems[26], incorporating this perspective in the theoretical framework offers valuable insight into how such it can be assessed from a quantitative point of view. However, while these metrics provide a structured way to compare outcomes, they do not ensure that they capture the complexity of user experience, contextual factors, or the nuanced impact of design decisions. From this perspective, we choose to present the previous assessment of efficiency

and effectiveness in the following sections 3.5.1.1 and 3.5.1.2 with an ambition of how its contributions together with other theoretical viewpoints could capture the complexity of the research questions of this study.

#### 3.5.1.1 Measuring efficiency

Efficiency can be assessed by measuring the time it takes for a chatbot to process and deliver a response [26]. By analyzing the interval between when the user's prompt is submitted and the system's response is delivered the researcher can gain a better understanding of the time it takes the user to get a response. Since the prompts can vary in complexity, the data can be categorized into groups of simple, intermediate and complex levels of prompt data to facilitate a greater oversight. Following this, the response time can be converted into a standardized score to facilitate comparison. The formula in Fig. 3.2 is used to calculate this score:

$$\text{Efficiency} = 100 - \left( \frac{\text{Actual Time} - \text{Min Time}}{\text{Max Time} - \text{Min Time}} \right)$$

Figure 3.2: Efficiency formula by Alabbas and Alomar. [26]

#### 3.5.1.2 Measuring effectiveness

By measuring the amount of incorrect responses provided by the AI, the designer can gain insight on how effective the user perceives the chatbot[26]. By enabling the user to rate the responses on a scale from very poor (1) to excellent (5), real time assessment of the provided answers can be measured. However, this requires well defined guidelines on how the scale should be interpreted by the user. The value can be calculated by first measuring how accurately the chatbot performs as shown in Fig. 3.3, and then normalizing the effectiveness score by converting it into a standardized 0100 scale and is shown in Fig. 3.4.

$$\text{Scaled Effectiveness} = (1 - \frac{\textit{Observed Error Score}}{\text{Maximum Error Score}}) \times 100$$

Figure 3.3: Scaled Effectiveness formula by Alabbas and Alomar. [26]

$$\text{Normalized Effectiveness Score} = \frac{(X - 1)}{(5 - 1)} \times 100$$

Figure 3.4: Normalised Effectiveness formula by Alabbas and Alomar. [26]

## 3.6 Effective Workflows in Natural Language Interfaces

Workflows can be defined as an organized collection of tasks designed to complete a business process. Tasks can be executed by software systems, individuals, teams, or a combination of these [28].

Cognitive load (CL) refers to the cognitive resources required to complete a task. The theory of CL was developed within the field of education but is applicable to a variety of contexts, including user experience of graphical interfaces. CL can be divided into three components, intrinsic load, germane load, and extraneous load. Intrinsic load is connected to the complexity of the task or a system and is determined by the individual's cognitive resources. Germane load relates to the process of identifying and learning patterns in a task, within the research of CL there is a divided opinion if germane load should be considered as a part of intrinsic load or as a separate component. Extraneous load depends on the presentation and design of the interface. Extraneous cognitive load can be influenced by design choices and serves as a key consideration in human-computer interaction design. [29]

In the context of completing tasks through a natural language interface (NLI) Do et al. [30] highlights abstraction matching as a significant challenge, referring to the difficulty of composing a prompt that matches the system's capabilities. Continuous failures of abstraction matching can contribute to user frustration and impact the users technology acceptance and possibly lead to abandonment of the system. [30]

To approach this challenge, Do et al. proposes a set of conversational interfaces supporting grounded abstraction matching by applying the principle of least collaborative effort in communication grounding to the design. The study compared three variations of grounding interfaces with an ungrounded control interface. The set of grounded interfaces consisted of one conversational grounding interface, a multiple grounding interface, and a structured grounding interface. Do et al. recommends designing natural language interfaces that support provisional inputs and enable collaborative refinement between the user and the system, following the principle of least collaborative effort. Their findings showed that grounding interfaces, especially the ones offering structured input fields, can reduce users cognitive load, improve task performance, and increase system acceptance, still avoiding users feeling constrained. Furthermore, the authors propose that, for goal oriented natural language systems, using an structured guidance approach is more beneficial than aiming for fully free form naturalness. The structured support showed to help users effectively compose inputs while maintaining their sense of control. [30]

# 4

# Methodology

This chapter introduces various methodologies for collection, analysis, and validation of data. The chapter begins with presenting the methods used for data collection, including literature study, questionnaire, and interviews. This is followed by a section describing thematic analysis, log data analysis, personas and user journey mapping. The chapter concludes with describing the Delphi method, used as a validation method.

## 4.1 Data Collection

### 4.1.1 Literature Study

A literature study is conducted to identify the research gap, and gather insights from previous research including existing theoretical models and frameworks, serving as a basis for conducting new research [31, p. 131], [32]. The practical approach of a literature study involves scanning documents to provide an understanding of existing literature in the field, including identifying key topics. By identifying relevant themes and topics, sources and concepts can be organized accordingly to form the structure of the literature study. The writing process can be initiated once a general outline has been formed [32].

### 4.1.2 Questionnaire

Questionnaires are an effective method to gather a large amount of data and generate statistical insights about the problem area. It is a flexible approach, and often more efficient than interviews as it enables gathering quantitative data from a broader group of respondents. Furthermore, the direct interaction between the researcher and respondent is minimal, reducing the risk of bias and the researcher affecting the answers. [33, p. 95] However, the design of the questionnaire is crucial to ensure reliable and unbiased answers. Since the respondents don't have the opportunity to ask follow up questions or ask for clearer explanations of the questions it can lead to misunderstandings and the miss interpretations of the questions. The structure of the questionnaire can also affect the validity of the responses; if the respondent has access to all questions from start, it could adjust the answers to align with certain experience patterns, affecting the research validity and decrease the ability to study variables independently. [33, p. 95] Aiming to gather a broad range of perspectives

from the target group, to inform further data collection, a questionnaire is well suited for this study to gather initial insights.

### 4.1.3 Qualitative Interviews

Conducting qualitative interviews is a data collection approach that allows for an open and explorative approach to capture a detailed understanding of the problem and nuanced perspectives. In addition to the participants' verbal response, interviews can reveal valuable insights through the respondents tone of voice, as well as facial expressions and pauses answering the questions unlike questionnaires where there is limited interaction between the researcher and the respondent. [33, p. 104105], [34, p. 189]

Testing the questions in preparation for the interviews can provide insights about how the interview questions are working, and is a chance to improve the questions before the actual interviews. The researchers also get a chance to practice asking questions, and taking notes of the answers [34, p. 203]. Conducting interviews supports the collection of more detailed and nuanced information, complementing the broader insights gained through questionnaires. This method enables a deeper exploration of individual experiences and perspectives, which is essential for addressing the research questions in greater depth. As an alternative approach to interviews, observations can be utilized. By observing users in their natural settings, the observers can gain insights on actual user behaviours rather than how they describe their actions through interviews [35]. This approach could benefit this research as an alternative approach as well as a compliment if sufficient time is available.

## 4.2 Analysis

### 4.2.1 Thematic Analysis

Thematic analysis is a method used to identify, analyze, and interpret patterns forming themes in qualitative data. The method is commonly used to explore and capture respondents' experiences, beliefs, and behaviors from the data [36]. The practical approach for thematic analysis includes six phases as shown in Fig. 4.1. The *initial phase* focuses on familiarization with the data, to gain a deeper understanding about the content and identify potential themes. In*phase two* initial codes are generated about interesting elements in the data, working as foundational building blocks for further analysis. *Phase three* revolves around searching for themes through analyzing similarities and differences in the codes, and grouping them in broader themes. In *phase four* the identified themes are evaluated by comparing them to the original data on two levels. The first level aims to control that each theme is well founded in the coded data. The second level ensures that the themes are well founded in the full data material. *Phase five* includes defining and naming themes, the aim is to clearly describe the meaning of each theme and what is of interest about them, as well as how different themes are related to each other. In the *final phase* the analysis is presented in a report, including citations from the data to illustrate

the identified themes [37]. Aiming to identify key factors, thematic analysis serves as a valuable approach for revealing such factors through the systematic interpretation of themes emerging from the collected data.

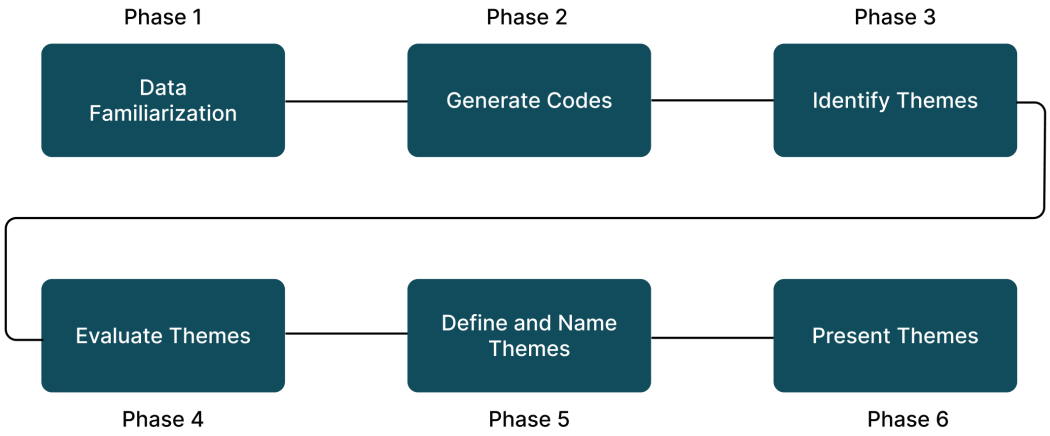


Figure 4.1: Visualization of the thematic analysis process described by Braun and Clarke.[37]

### 4.2.2 Log Data Analysis

Log data analysis is a method that can be utilized to analyse users' behavioural interaction with digital systems through logged data [38, pp. 349–369]. The method is often used in HCI research and requires behavioral logs that capture a large variety of events where users interact with the system. Dumais et al. describes large scale logs as recordings of events that are captured in real-world environments. Since the data is conducted in natural settings, it can be viewed as natural observations where the user is not influenced by the experimenters or the observers. Log data analysis allows researchers to form an abstract perspective of a substantial number of users and their behaviours while also enabling identification of usage patterns .[38]

Following the approach outlined by Dumais et al., Log data analysis is conducted over three phases and visualized in Fig. 4.2.

> The *initial phase* is *Data collection*, where the researchers find useful logs that capture what queries users tend to issue.
>
> The *second phase* focuses on *Data cleaning*, where the researchers initially familiarize themselves with the data before removing duplicates, filtering out irrelevant data, and anonymizing parts if necessary.
>
> The *third phase* is dedicated to *Using log data responsibly*, meaning that the researchers process and analyze the prepared data with careful consideration of user privacy [38].

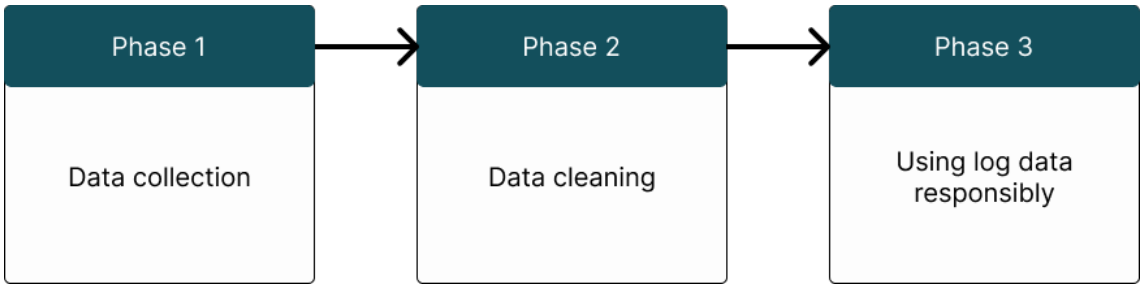


Figure 4.2: Visualization of the Data log analysis process by Dumais et al. [38]. Illustration by the authors.

Although log data analysis is commonly applied to quantitative research [39], it offers a strong potential as a part of a mixed methods approach alongside thematic analysis in this study. When analyzing data sets provided by Recorded Future, log-data analysis could serve as a way of observing behaviors for a diverse group of users while investigating common patterns in successful and unsuccessful user experiences.

### 4.2.3 Personas

Personas are a representation of shared characteristics and properties among multiple target groups and are commonly used for creating a better understanding of the user within the design team [40]. This methodology allows for the analysis of both quantitative and qualitative data, ensuring that the representation is well grounded in real user behaviors and needs [41], [42].

Given the potential to gain valuable insights from existing user data, Quantitative Persona Creation (QPC) presents an opportunity to enhance efficiency in persona development [43]. While QPC allows for a large scale, data-driven approach, our goal remains to integrate both qualitative and quantitative insights to achieve a more comprehensive and in depth perspective on user needs.

### 4.2.4 User Journey Mapping

User journey mapping is widely used within the field of UX design to capture how a user interacts with a product or service, visualizing the step-by-step experience. The aim is to map the different phases of the users journey from planning to completion using a specific product or service. A common approach is to visualize the phases on a horizontal axis to clarify the time progression. On the vertical axis different metrics of interest can be added.

User journey mapping can serve as a complement to personas. While personas focus on a static view of a typical user, user journey maps cover a dynamic and time based description of the users experience, adding a valuable 3rd dimension. [44], [45]

## 4.3 Validation

### 4.3.1 Delphi method

The Delphi method is a well established approach for evaluating a subject through the collection and analysis of expert opinions [46]. The method offers a potential means for validating various outcomes, such as derived guidelines with the assessment of experts within relevant fields. In this study, the Delphi method may contribute by offering an additional means of triangulating the forthcoming findings.

The method consists of two main segments,*selecting experts* and, *conducting expert advice* and can be viewed in Fig. 4.3. Selecting experts is considered a critical process in the Delphi method where the qualified experts need to have a deep understanding of the issue. To facilitate this, 5 additional steps are used to select relevant experts using a Knowledge Resource Nomination Worksheet (KRNW).

*Step one* focuses on preparing the KRNW by categorizing appropriate classes of experts before identifying them individually. It is essential to ensure that all relevant experts are considered to have the opportunity of diverse insights on the issue. This is done by preparing the KRNW with the categories: organizations, disciplines or skills and related literature.

*Step two* involves populating each identified category with potential experts who possess significant knowledge of the field of study. This process helps create a clearer overview of which individuals are likely to contribute valuable insights and assessments.

*Step three* allows for the expansion of the expert list. At this stage, the study and its objectives are defined more clearly to determine whether additional expertise is needed. As researchers, we may also contact already identified experts for further recommendations of individuals who may be well suited to the studys requirements.

*Step four* focuses on prioritizing and ranking the experts based on their qualifications. By organizing them into sub-groups, they can be ranked according to criteria such as years of experience, publications, and geographical relevance. It is acceptable at this stage for experts to appear in more than one list if they possess overlapping qualifications.

*Step five* involves inviting the highest-ranked experts to participate in the study. It is recommended that 10 to 18 experts be included in the next segment. Each invited participant should be provided with clear information about the studys objectives, timeline, expected value, and the estimated time commitment required for their involvement.

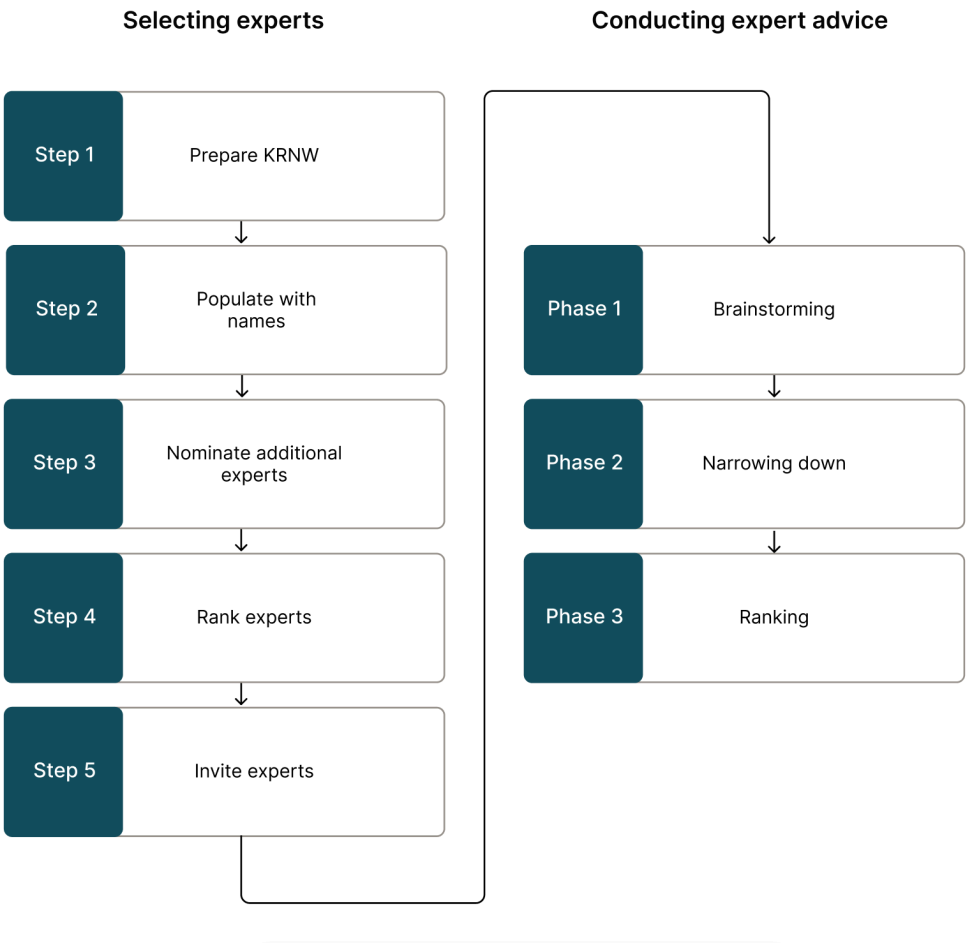


Figure 4.3: Delphi method process after Okoli and Pawlowski. [46]

Conducting the experts advice is carried out through three phases (1) Brainstorming, (2) Narrowing Down Factors and (3) Ranking factors.

*The brainstorming phase* consists of two short questionnaires. Based on Okoli and Pawlowski, the first questionnaire commonly recommends each expert to independently list and rank factors they consider important together with a brief explanation for each decision [46]. In this study, the goal is to transform our findings into practical guidelines that can assist the development of AI-driven solutions for B2B consumers. The first questionnaire would therefore act as an individual assessment by each expert where they have the opportunity to list the guidelines together with their first impressions regarding relevance, inadequacies and remarks on any missing parts based on their expertise. After this, we as researchers consolidate the responses into a refined list of factors. In the second questionnaire, experts review and validate this summary before moving on to the next phase.

*The narrowing down phase* focuses on identifying the most important

guidelines and factors from the consolidated list. The experts are grouped into panels based on their stakeholder roles to allow viewpoints from experts with similar priorities to align. In this phase each panel prioritizes the top 10 factors and guidelines they consider significant.

*The ranking phase* focuses on having each panel establish a prioritized order of the most important factors and guidelines. Each panel works independently and rank them based on their perceived significance.[46]

The application of the Delphi method could serve as a valuable approach for validating the findings that have emerged from this study. By incorporating expert perspectives, valuable insights may provide an indication of confirmation or alternatively, point to additional considerations that would be needed in further refinement. Online Expert Panels could serve as an alternative approach if physical interviews are not possible to achieve using the Delphi Method. Online Expert Panels enables the inclusion of a large number of stakeholders across different geographic locations and would therefore facilitate a bigger diversity of experts [47].

# 5

# Process

This chapter describes the execution of the research. The chapter begins by presenting the process of defining the research problem. This is followed by describing how each of the methodologies used for data collection and analysis were utilized. Additionally, the findings that informed the final results are presented under each methodology. The chapter concludes with outlining the process of creation and validation of design guidelines. The execution process of this research is described below in Fig.5.1.

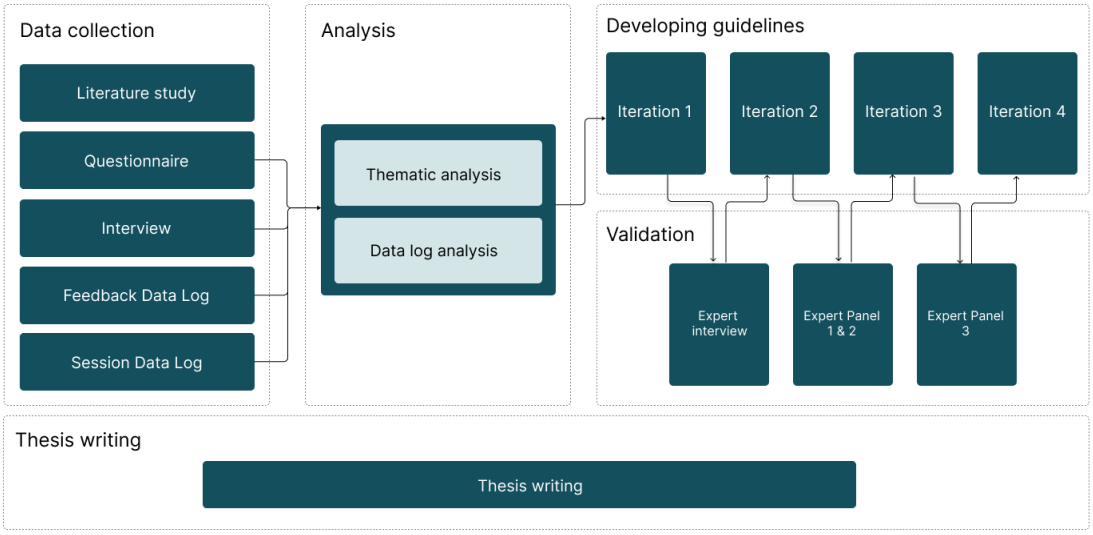


Figure 5.1: Execution of the research process.

## 5.1 Define Problem

The study was initiated with two informal meetings with the supervisor at Recorded Future, the aim was to form a deeper understanding of the research field, current challenges, and possible perspectives the challenges could be investigated through. Simultaneously, exploratory research was conducted to further understand the current research done in the field. A first definition of the research field was created, focusing on the role of UX design to improve complex knowledge information systems. After more in depth discussions with our supervisor, the focus was narrowed

down to specifically target the role of UX design in AI-driven complex knowledge information systems.

## 5.2 Literature Study

Once the scope of the study was defined, a literature study was conducted to build a strong theoretical foundation focusing on two main aspects; literature related to domain knowledge and design related literature. The process involved searching in academic databases. The main search engines used for our search were Chalmers Library's database and Google Scholar. In addition, we reviewed papers recommended by our supervisor at Recorded Future, which helped introduce us to key works within the field. Search terms such as UX design", usability, AI, natural language interfaces (NLI) and related concepts were used to locate relevant literature. These keywords reflect the central focus of our research and aims at exploring how users interact with AI systems, particularly through natural language.

The selection process emphasized peer-reviewed publications to ensure the credibility and academic quality of the literature. We primarily searched for research and theoretical frameworks published in journals and conferences within the field of human-computer interaction (HCI), user experience (UX), and artificial intelligence (AI). When selecting which studies to include, we considered how well each source related to our research question. We aimed for a combination of well established studies and more recent published papers.

## 5.3 Questionnaire

In this study, the survey data used were collected as part of an internal research project conducted by Recorded Future. A total of 39 participants responded to the survey. All participants in the survey work at companies that are customers of Recorded Future, and use the modules where the AI functionality is available. The questionnaire was designed and distributed by the AI development team and the selection of participants was also carried out independently of the authors. As part of this masters thesis, we were granted access to the anonymized survey responses for the purpose of answering the research questions. While we did not contribute to the development of the questionnaire or the recruitment of respondents, our role involved conducting a thematic analysis of the qualitative data in order to identify patterns and themes related to the platform's AI reporting functionality. The survey was distributed with the intent of examining several aspects of users current use of Recorded Futures AI. Several of the questions covered topics that fell outside the scope for this thesis, which included users general product habits and work roles. However three of the questions were included in the study as they captured the users experience and perceptions of the AI functionality making them highly relevant for answering the research questions.

The thematic analysis process was conducted according to the six steps outlined by Clarke and Braun [37]. The process was initiated with a familiarization phase,

involving reading all responses multiple times. Initial codes were then generated inductively from the data using manual coding. These codes were iteratively reviewed and organized into candidate themes, which were refined through multiple rounds of analysis to ensure clear differentiation. Each theme was evaluated to ensure alignment with both the coded data and the original data material. Each theme was clearly defined and named to capture the essence of the participants perspectives. To complete the analysis process an overview was created presenting the themes with descriptions and citations as support.

### 5.3.1 Findings from Questionnaire

This section presents the user experience survey results, focusing on the users general experiences and perceptions of the first six months with the AI reporting function. A total of 39 participants responded to the survey.

The result from the thematic analysis highlights several key themes that influence how diverse professional users interact with and experience a natural language interface for AI-driven report generation. The following five themes were identified in the analysis: *Customization and Flexibility, Efficiency and Effectiveness, Workflow Considerations, Transparency and Trust*, and *Language Model Performance*.

> *Customization and Flexibility* concerns the systems ability to support diverse user needs, including how well it can be adapted to support individual use cases and preferences.
>
> *Effectiveness and Efficiency* reflects the users perception of how well the system supports meaningful value creation, as well as how quickly it enables them to achieve their intended goals.
>
> *Workflow Considerations* relate to how the system aligns with existing organizational processes and desired outcomes.
>
> *Transparency and Trust* addresses the users ability to understand how outputs are generated and the sources of the information.
>
> *Language Model Performance* refers to how well the systems language based features align with the users expectations in terms of relevance, accuracy, and the level of detail provided in the output.

The majority of the responses related to *Customization and Flexibility* focused on the level of detail in the generated output. The ability to adapt the timeframe of the output also emerged as a recurring feedback within this theme. Feedback within the theme *Effectiveness and Efficiency* revolved around users being uncertain about their prompt writing skills, as they couldnt achieve their intended goal. Furthermore, some users highlighted efficiency challenges as the responses required finetuning and that the system offered a limited contribution in relief of workload. Within the theme *Workflow Considerations* the feedback stated that the current functionality is suitable for specific use cases. Feedback regarding *Transparency and Trust* related to users struggling to evaluate the reliability of the information or experienced lack of citations. Several of the responses related to *Language Model Performance* indicated

dissatisfaction with the level of detail in the response, and a few users experienced that their instructions were not reflected in the output.

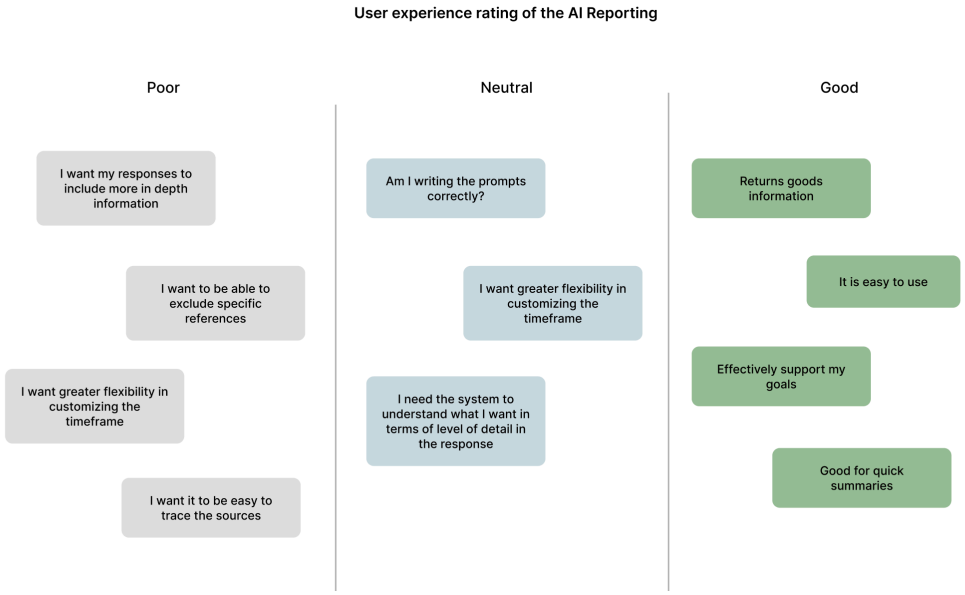


Figure 5.2: User experience rating of the AI Reporting, showing qualitative feedback related to a specific rating. The citations are not direct quotes from survey responses, but rather anonymized and rewritten as statements to exemplify the nature of feedback from each category.

The majority of the respondents rated the function neutral. A few of the users expressed uncertainty if they were asking the correct questions to achieve their goal. Recurring feedback in this section also highlighted the desire for more customization of the output, along with uncertainty about the reliability. Users that have had a good experience with the functionality reported that the system is very easy to use and returns good information, effectively supporting the needs of their specific use cases. Further, they highlighted that the system meets their needs well regarding providing good summaries quickly. Users that reported a poor experience with the system often experienced issues regarding achieving the correct level of detail in the response, looking for greater depth in the response. They also expressed the need for more customization options and flexibility regarding sources and timeframe settings. Lastly, the need for easier validation of reliability and traceability of the sources was communicated.

The survey also captured specific user needs regarding affordances within the system. The majority of the reported needs was related to customization of the visual design of the report, as well as allowing for a more flexible and iterative process regarding selection of sources. Furthermore, support for formulating natural language requests was reported in the responses.

## 5.4 Interviews

In line with the approach for the survey data, the interview data used were collected as part of an internal research project conducted by Recorded Future. The interview guide was designed by the AI development team and the selection of participants was limited to the respondents from the survey that had given consent to be contacted for setting up an interview. The selection process was carried out independently of the authors. As part of this masters thesis, we were granted access to observe one of the interviews and watch a recorded video of a second interview. Twelve of the survey respondents agreed to be contacted for setting up an interview, our aim was to take part in 6-10 interviews. However, during the weeks working with data collection only two interviews were scheduled. The interviews acted as a support for the survey analysis results, in adding more depth and detail to the insights. As an initial step of the analysis process, the interviews were transcribed. Thereafter citations from the interviews were grouped under themes identified from the survey analysis.

### 5.4.1 Findings from Interviews

Two of the respondents from the survey participated in interviews. The interviews resulted in collected data capturing detailed insights about their user experience with the AI reporting function.

The themes *Transparency and Trust and Usability and User experience* were central topics of the interviews. To enhance *Transparency and Trust* users want to be able to add their own files and sources to generate the report from. Furthermore, users want to be able to trace if the source is the original. Regarding *Effectiveness and Efficiency*, customization of the report design that allows for a more iterative workflow is considered to be important. Generating reports with visualizations was also a central feedback of the interviews.

## 5.5 Analysis of User-AI Interaction

To gain a deeper understanding of the interaction between users and the AI, an analysis was conducted on feedback data and session data derived from past sessions. The data used in this analysis covers a six month period and was obtained from existing logs provided by Recorded Future. All datasets had been anonymized by the company prior to the analysis in order to ensure user privacy. Although we did not collect the data ourselves, the dataset was extensive and required a procedure of processing and extracting relevant information. Our work involved identifying, filtering, and organizing the data in a way that enabled meaningful analysis aligning with the goals of the study. The analysis was grounded in central topics identified in the literature study; *affordances*, *trust*, and *effectiveness* in the AI system. The feedback data was utilized as a delimitation, thereby enabling a distinction between successful and unsuccessful sessions between the user and AI. The session data, containing complete conversations between users and the AI enabled a deeper analysis

of how different scenarios can unfold. The following sections will provide a more detailed explanation of how the execution of these two analyses were carried out.

### 5.5.1 Feedback Data Analysis

The user feedback data was first organised to distinguish successful and unsuccessful conversations between the user and AI. The dataset contained three different categories of user feedback; *thumbs up* indicating positive feedback, *thumbs down* indicating negative feedback and additionally a possibility to leave a *comment* together with the thumbs down feedback. Each user feedback was directly linked to a specific response written by the AI, and with an association to the session the message was retrieved from.

The initial step was to sort and divide all the collected data into the three categories; Positive feedback, negative feedback and negative feedback with comment. Due to the valuable depth of qualitative data, a thematic analysis was conducted on the category negative feedback with comment to identify recurring themes that could provide insights on lacking or missing elements of the NLI. The process followed Clarke and Brauns six-step thematic analysis [37] and began with familiarization of the dataset. Due to the large dataset of 192 rows, we chose to carefully skim through all the data before moving on to individually code each feedback comment. Themes were developed based on these codes and guided by patterns of similarity while also ensuring that each theme accurately represented the data it was intended to capture. After this, the themes were reviewed and refined to ensure clear formulations and descriptive names. In the final phase, representative quotes were selected to support and illustrate each theme. To summarize how prevalent each theme was within the dataset, we calculated how many comments were linked to each theme.

#### 5.5.1.1 Findnings from User Feedback Data Analysis

The following section presents the result of a thematic analysis based on user feedback from conversations with AI generated responses. The feedback is in the form of a users negative reaction to a message in a conversation, with an additional comment with qualitative feedback. The analysis includes 192 feedback comments. The coded citations were initially categorized into 14 distinct themes. Five of these were excluded from further analysis, as they were closely related to technical aspects of the system and fell outside the scope of this thesis. The remaining nine themes were then grouped into four main themes, as several of them were closely related and shared overlapping content. The analysis resulted in four themes: *Miscommunication, Level of Detail Issues, Trust Issues,* and *False Affordances*.

> *Miscommunication* refers to when the AI doesnt understand the users intention. The user perceives that the AI does not follow their instructions, the user is confused by the response, or the user is frustrated as a result of trying to achieve something through the conversation.
>
> *Level of Detail Issues* is related to when the system fails to reflect the users needs in terms of level of detail in the response.

*Trust Issues* refers to the situations where the user can't verify the source of the information or the user questions the reliability of the information.

*False Affordances* captures the situations where the user is trying to get the system to do something that is not possible.

### 5.5.2 Session Data Analysis

To further deepen the insight gained from the feedback data, the session data was used to provide a more detailed understanding of the context behind the users feedback. This was achieved by analyzing entire sessions that include all user prompts and corresponding AI responses in the same chat. A mixed method approach was employed, combining log data analysis and thematic analysis to examine user interactions.

The process was initiated using the log data analysis's three phases [38] where the first step was to collect a comprehensive dataset (*Data collection*). We obtained six months of session data containing conversations between the user and the AI. The second phase consisted of data cleaning, during which private and irrelevant data were removed. For the purpose of this study, we selected the conversations that had previously been identified in the feedback analysis as related to affordance, effectiveness, and trust for further analysis. In the third phase we conducted the analysis by *Using log data responsibly* and incorporating valuable steps from the thematic analysis to detect common patterns. Each session was reviewed while key codes were identified and subsequently analyzed for common patterns. Additionally, each session was evaluated to determine whether the user had attempted to use fine-tuning. This was to get an indication of how the user intended to guide the AI in cases of potential misunderstanding. In regards to using the log data responsibly, the analysis was conducted entirely on our local computers. The final results were later rephrased to ensure that no sensitive data would be exposed or disclosed to unintended recipients.

#### 5.5.2.1 Findings from Session Data Analysis

To get a deeper understanding of the user's experience, 88 full AI sessions connected to the user feedback regarding each separate theme were analyzed with a mixed-method approach, including log data analysis and thematic analysis. From this point forward, AI sessions will also be referred to as *conversations* throughout the report.

The analysis for the theme Miscommunication includes 50 conversations. The analysis resulted in five patterns that characterize miscommunications in an AI conversation. The following patterns were identified: *Incorrect Scope, Incorrect Timeframe, Incorrect Level of Detail, Abstraction Mismatch,* and *Missing Contextual Carryover.*

As represented in Fig. 5.3 the result shows that 45.5% of the use cases connected to miscommunication were related to the AI giving a response with out of scope information. Several use cases include examples where the AI only focuses on specific parts of the prompt, not capturing the full scope. Furthermore, there are use cases

showing that the AI misunderstood the users question, leading to an out of scope answer as irrelevant information is presented. 27.3% of the use cases exemplifies situations where the AI does not apply the requested time frame to the response. Additionally, for the use cases related to creating a report the user tries to prompt a change of timeframe outside of the default options provided in the report template editor. 15.9% of the conversations related to miscommunication can be connected to incorrect level of detail in the response. In most use cases the user expresses a need for a more detailed response to be able to gain value from the response. 6.8% of the use cases were related to the user framing the prompt at a different conceptual level than the AI is designed to handle leading to an abstraction mismatch, where the AI couldnt find data matching the query. Lastly, 4.5% of the use cases involved the AI missing contextual carryover, where the AI failed to incorporate information from earlier messages and indicated a lack of contextual memory or thread awareness.

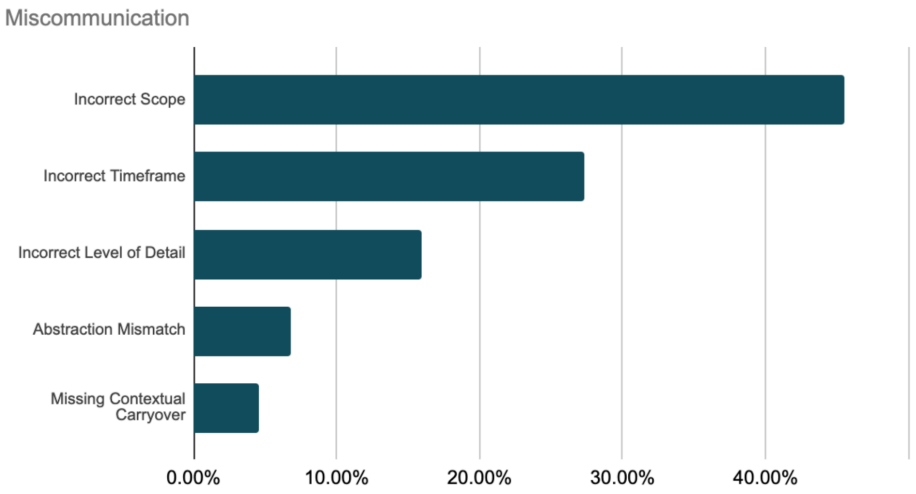


Figure 5.3: Results showing underlying patterns connected to miscommunications.

The analysis for the theme *Level of Detail Issues* includes ten conversations. The analysis resulted in three situations that can be traced to affect the user experience negatively due to the level of detail in the AI response. The most common situation is when the response is too generic to match the user's intention with the prompt request, such as missing proper explanations or the response reflects what the user considers to be a summary when seeking more in depth information. Another situation is when the response is not considered to complete, lacking thoroughness. Lastly, is a situation when the AI provides details in the response that the user considers to be irrelevant. In one of the ten examples the user tried to fine tune the answer to achieve a more detailed response.

The analysis for the theme *Trust Issues* includes 15 conversations. The analysis resulted in four identified situations that can affect the trust negatively with the AI. The following patterns were identified: *Traceability Issues, Unreasonable Response, Outdated Information, and Hallucinations.*

In most of the situations, the lack of possibility for the user to trace the source of

the information was causing trust issues. Another recurring situation was that the user considered the response to be unreasonable, due to the intended meaning of the source not being fully captured and therefore lacking crucial information. A third situation causing trust issues was when the user perceived that outdated information was included in the response, therefore user uncertainty about accuracy led to trust issues. Lastly, use cases when the user suspects hallucinations led to trust issues, meaning the AI providing made up information or details due to lack of data related to the prompt request.

The analysis for the theme *False Affordances* includes 13 conversations. The analysis resulted in two situations that reflect false affordances within the AI. In twelve of the use cases, the user wanted to get information within a certain timeframe but couldnt get the AI to apply the asked timeframe. In one use case the user wanted the AI to collect the data from a certain module in the platform, which is not a possible action.

## 5.6 Creation of Guidelines

The process of creating the guidelines began by listing the factors identified as important for enhancing the user experience of a NLI. Each factor was then linked to one or more of the reports three central theoretical topics: *affordances*, *trust*, and *effectiveness*. These connections were further developed into one or more guidelines, based on the findings from the results combined with the theoretical foundation established through the literature study. The initial list of guidelines added up to a number of 13 recommendations.

### 5.6.1 Mapping Identified Factors under Affordance, Trust, and Effectiveness

This section presents the identified factors in a categorized table 5.1, following the division of the three focus areas that were identified in the literature study: *Affordance, Trust,* and *Effectiveness.* This division is made to ensure that the upcoming design guidelines address the important factors in our findings and facilitate the translation of user needs and existing challenges into actionable design guidelines. The marked boxes next to each factor indicate a required improvement within the area in order to address the associated challenges. By employing the insights from the literature study and the patterns found in real world user experiences, this work outlines a foundation for formulating design guidelines.

*Customization and Flexibility* reflect challenges related to both affordance and effectiveness, as the findings show that NLIs lacking these qualities often fail to support users in maintaining an effective workflow or task completion. This is frequently linked to users not understanding the system's capabilities and limitations which can indicate *False Affordances.*

*Effectiveness and Efficiency* issues were commonly seen throughout the survey, conversation and feedback analysis where users struggled to carry out their tasks and

calls for an improvement regarding effectiveness. This also includes how quickly it enabled them to achieve their intended goals.

*Workflow Considerations* can be connected to affordance and effectiveness challenges as feedback indicated that the system supports some of their use cases but not all of them.

*Transparency and trust* was identified as one of the most recurring factors among the conducted analysis. Users clearly state that they want to be able to trace the sources used for generating the output, as a way to validate the reliability of the response.

*Language Model Performance* is a factor related to both trust and effectiveness, challenges concern relevance and accuracy of output where the user wants to both be able to trust the response as well as achieving alignment between their intention with an input and the related output.

*Miscommunication* reflects challenges related to all the focus areas. The reasons identified leading to miscommunications were users trying to achieve false affordances, or struggling to achieve their goals reflecting effectivity challenges. These situations relate to trust issues when the system continuously fails to interpret inputs as intended.

*Level of Detail* reflects effectiveness challenges as users experience issues with achieving their goal in terms of level of detail in the response.

*False Affordance* reflects affordance challenges, specifically regarding situations where the user thinks an unsupported action is possible.

Table 5.1: Factors mapped to focus areas; *Affordance, Trust,* and *Effectiveness.*

| Factor | Affordance | Trust | Effectiveness |
|---|---|---|---|
| Customization and Flexibility | X | | X |
| Effectiveness and efficiency | | | X |
| Workflow Considerations | X | | X |
| Transparency and Trust | | X | |
| Language Model Performance | | X | X |
| Miscommunication | X | X | X |
| Level of Detail | | | X |
| False Affordances | X | | |

#### 5.6.1.1 Iteration 1 - Presentation of Initial Guidelines

To initiate the development of the Design guidelines, a first draft was created based on the key factors and challenges identified through data analysis and insights from existing research. The result from the first iteration is presented in Fig. 5.4 below.

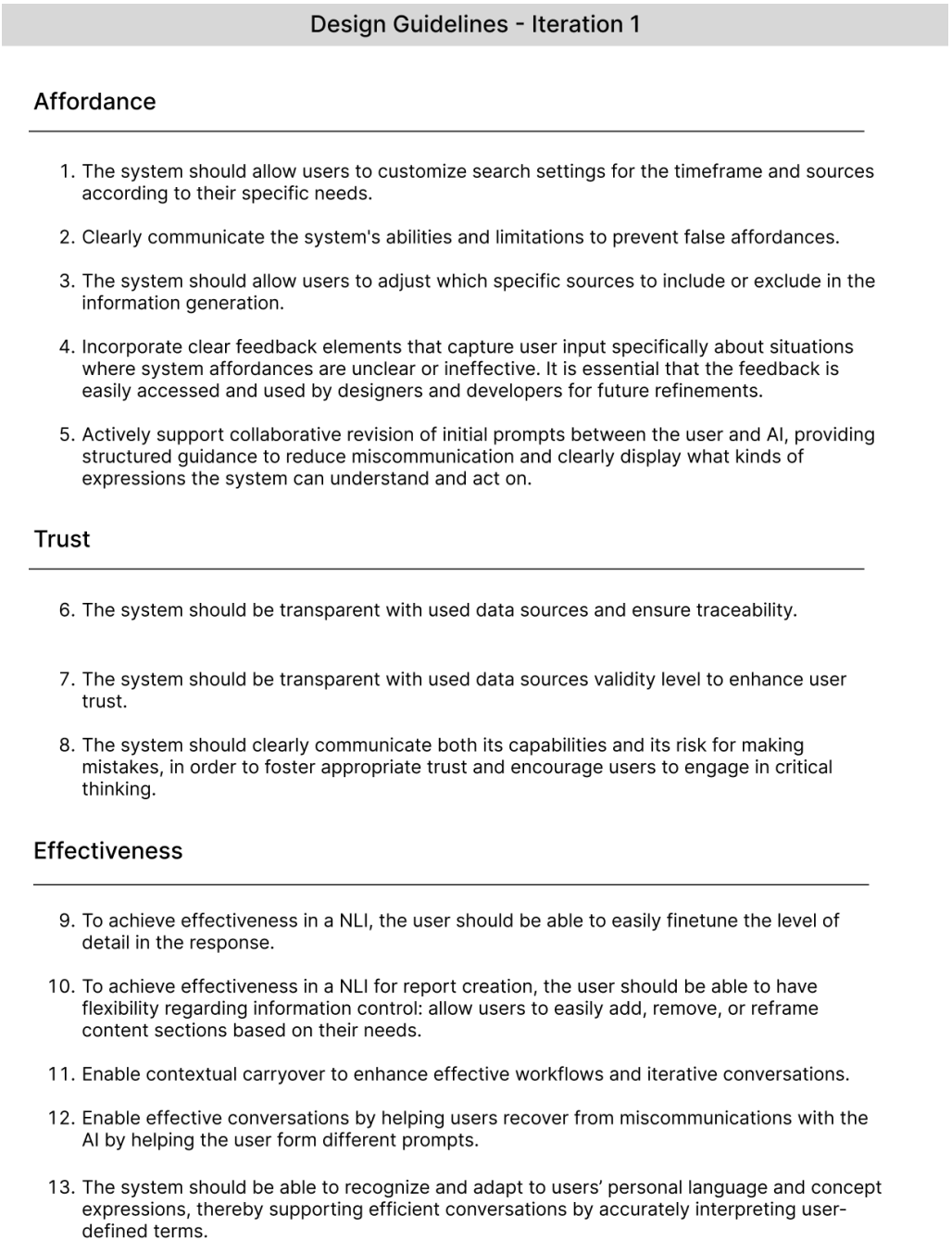


Figure 5.4: Initial Guidelines.

An expert interview and expert panel evaluations were then conducted to validate the guidelines over three additional iterations. The following section provides a detailed description of the expert validation process.

## 5.7 Expert Validation

The expert validation was carried out to assess and refine the formulated guidelines, following the process of the Delphi method. As stated in section 4.3.1, the method consists of two segments, *Selecting experts* which involves five additional steps and *Conduction expert advice*, consisting of three additional phases. The execution of this process ultimately led to an alternative approach for the second segment, which is presented in the following section.

The validation started with preparing the KRNW. In the first step, three appropriate classes of experts were selected, Experts within AI development, UX research and UX design. In the second step, these classes were populated with 12 potential experts whom we considered relevant for the purpose of the study. In the third step, the list was further reviewed and expanded to 16 potential experts in collaboration with one of the experts. Step four focused on ranking the experts. This was done by evaluating them based on their years of experience within the discipline and level of involvement in this project. With the aim of involving at least 12 participants in the study, 16 experts were invited during step five of the expert selection phase. This approach was intended to ensure a sufficient number of participants, accounting for the possibility that some experts might be unable to participate due to conflicting tasks or time constraints.

The second segment, *Conduction expert advice*, was originally designed to allow the experts to brainstorm potential improvements and rank the guidelines through two iterations. Due to concerns regarding a potentially low response rate to the questionnaire, an alternative approach was adopted and presented in Fig.5.5. This allowed for a more flexible arrangement with the experts while still capturing a diverse range of feedback. Moreover, the revised approach facilitated multiple iterative cycles, which in turn provided additional time for the guidelines to evolve and mature into a final version.

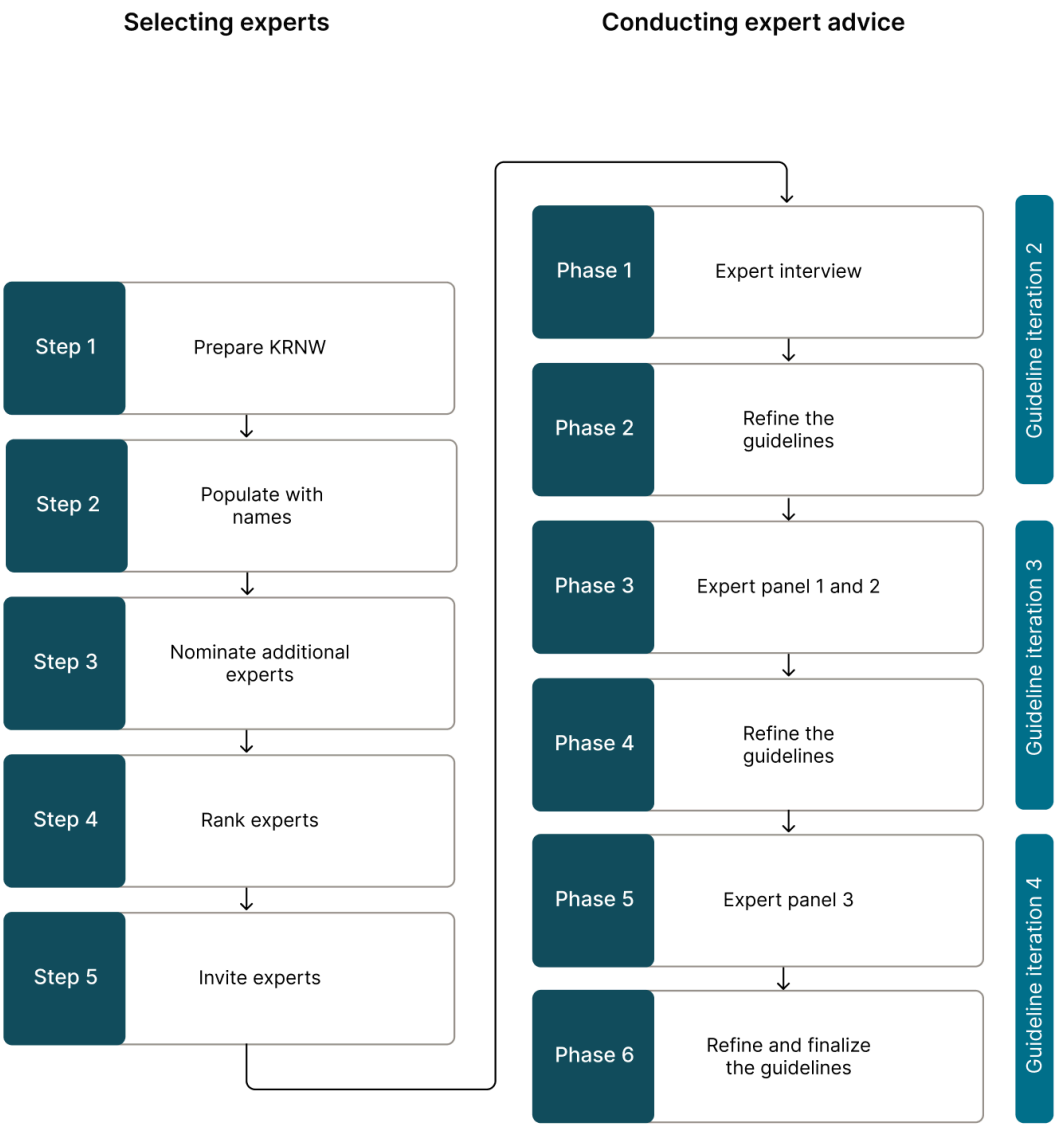


Figure 5.5: An adjusted approach to the Delphi method, based on the framework by Okoli and Pawlowski [46]

*Phase 1*: Instead of utilizing two questionnaires as a second iteration of the guidelines, we asked our top ranked expert within the UX research discipline to sit down to evaluate the formulated guidelines. The expert was provided with the list of guidelines in advance to review prior to the first assessment. The following day, the expert provided feedback during an interview, in which all proposed changes were documented collaboratively to ensure accurate interpretation.

*Phase 2*: In the second phase, the guidelines were adjusted based on the received input. The initial list of guidelines consisted of 13 potential guidelines. After applying the collected feedback onto the guidelines, the updated list had four rewritten guide-

lines, one new added and one removed. Following that, the guidelines were placed in a document and sent out together with an updated invitation to the selected 15 experts.

*Phase 3*: In the third phase of the alternative approach, two expert panels were invited to review and give feedback on the refined guidelines. The phase was carried out during two separate meetings where each panel discussed and left comments in a common document as we observed the discussions. The panels were divided based on their disciplinary background to capture a broad range of perspectives and domain specific insights. However, due to conflicting tasks within the UX Research and UX Design team they were grouped together and viewed as a UX focused panel, based on the number of individuals involved. This resulted in the AI developer panel consisting of seven participants and the ux panel consisting of three participants.

*Phase 4*: Following the collection of expert feedback the guidelines we once again refined. In this phase, we carefully analyzed the input to identify recurring suggestions and address valuable insights from the two panels. This process aimed to enhance the clarity, applicability and quality of the guidelines based on the received feedback.

*Phase 5*: In the final evaluation, a new panel of five experts within the UX discipline were invited to validate the refined guidelines. In this stage, the goal was to validate that the guidelines were well suited for practical guidance and if any final refinement were needed.

*Phase 6*: The guidelines were refined one last time based on the last expert panel.

### 5.7.1 Results from Expert Evaluations

This section presents the result from the conducted Delphi method evaluations. The following subsections are organized according to the individual expert assessments.

#### 5.7.1.1 Iteration 2 - Expert Interview

The following section presents feedback from an expert working with UX research.

The expert highlights that the guidelines target different levels, some more detailed and product specific while others are more open for interpretation. Regarding the category trust, the expert states that the guidelines capture important challenges addressed in literature. However, the expert pointed out that trust is an extensive challenge within this area and additional guidelines for this category could be developed.

On the category effectiveness, specifically for guideline 9 and 10, the expert elaborated on the challenge with managing ambiguity in the relationship between input-output size, emphasizing the problem becomes more severe the bigger discrepancy in this relationship. AI report creation is more of an ”order” than an iterative process, making it harder to refine and control the output.

#### 5.7.1.2 Iteration 3 - Expert Panel 1 and 2

The following section presents the feedback from expert panel 1, consisting of seven experts, working with AI development.

*Affordance Category Feedback* - For guideline 1, experts emphasized that the system should prioritize user inputs, especially in cases where the system affordances are unclear or ineffective. If interpretation is not possible, the system should offer customization options to guide the user. Regarding guideline 2, experts highlighted the importance of clear system feedback. They suggested that the system should recognize when a user attempts an unsupported expression and provide guidance, such as explaining why the input is not supported and offering alternative phrasing to achieve the desired outcome. Finally guideline 4, experts recommended using more natural language and avoiding overly technical terms. For example, replacing phrases like collaborative revision and "prompt" with more approachable alternatives like "guide" and "initial question". *Trust Category Feedback* - For guideline 7, experts recommended including the AIs reasoning process to enhance transparency, allowing users to understand how the AI arrived at its responses. Additionally, they suggested introducing a new guideline that acknowledges the varying language styles and characteristics of different models, which can significantly impact user experience. This guideline should also address the importance of understanding that not all models have the same level of contextual awareness. While a chat interface may present the interaction as a continuous conversation, the underlying AI may process each user input as a standalone request, which places different design requirements on the system. *Effectiveness Category Feedback* - Experts noted that the term "effectiveness" might be problematic as its definition varies significantly depending on the context and specific user needs. To truly assess effectiveness, designers must understand what their particular users consider effective in their workflows. Furthermore, they observed that the guidelines within this theme are mostly product specific, especially guideline 10, which may not apply universally across different AI systems. For guideline 9, experts suggested clarifying the perspective from which effectiveness is being measured. For guideline 11, they recommended using another term for "contextual carryover" and clearly defining this concept before enabling it, as its meaning can vary significantly depending on the extent to which it is applied.

*Additional Feedback* - Regarding the earlier mentioned comment on effectiveness being a challenging term for this category. They offered the perspective that if affordance and trust are addressed in the right way, it will likely lead to an overall effective system, as those factors form the foundation of intuitive and reliable systems. Additionally, they recommended clearly defining the scope of the NLI guidelines to ensure they align with the intended design goals and accurately capture the types of user interaction being prioritized.

The following section presents the feedback from expert panel 2, consisting of three experts, working with UX research or UX design.

*Affordance Category Feedback* - For guideline 1, the experts emphasized that designers have limited control over how the model itself functions. Instead focusing on what users are "allowed" to do, the guideline should prioritize clearly communicat-

ing the systems capabilities to users, aligning the interface with what the system can realistically support.

*Trust Category Feedback* - The experts questioned the distinction between guideline 6 and 7. Furthermore for guideline 6, they pointed out that different users assess source validity using different criteria, therefore recommending that the system should clearly explain what validity means within specific contexts. Furthermore, for guideline 7, they questioned whether the approach genuinely encourages critical thinking.

For guideline 8, they emphasize the importance of expressing the AI decision making in a human readable way to support users with limited technical knowledge and streamline the validation process.

*Effectiveness Category Feedback* - For guideline 11 on contextual carryover, the experts strongly recommend clarifying the scope. This could include a single session context, personal use history, or organizational context, each with different implications for data privacy and user expectations. Furthermore, they pointed out that not all users value this feature, so it should be clearly defined and appropriately targeted.

They also suggest elaborating on guideline 12, as its current phrasing was considered too vague to effectively guide design decisions. Additionally, they recommended adding a guideline focused on visualizations to increase effectiveness.

*Additional Comments* - From a designer perspective, the experts emphasized the need for more concrete examples within each guideline to provide clearer direction for designers, as the current phrasing is often too generic to be actionable.

Lastly, they highlighted that transparency alone may not always be beneficial, as users need the appropriate skills and time to interpret this information. Instead, guidelines should promote example based explanations that are easier for a wider range of users to understand.

#### 5.7.1.3 Iteration 4 - Expert Panel 3

The following section presents feedback from expert panel 3, consisting of five experts working with UX design or UX research. This group was presented with two versions of the guidelines. The guidelines presented to expert panel 1 and 2, as well as an updated version of the guidelines based on the feedback from iteration 3.

Regarding clarity and readability of the guidelines, the experts think that the guidelines are written in an overly academic style, requiring multiple readings to fully understand. Many sentences are long and cover multiple guidelines, making it difficult for designers to assess their application. Additionally the experts suggest simplifying the language to ensure designers with varying levels of AI knowledge can follow the guidelines. They encouraged us to reduce unnecessary wording. Consider how much text is necessary to describe the core message without losing meaning. Furthermore, they suggest to differentiate between guidelines covering model affordance and UI affordance to add clarity. Comments regarding terminology and definitions revolve

around that many terms used are too general and open to interpretation. They suggest we clearly define key terms, like ”affordance” to avoid misunderstanding. Referring to the structure and organization of the guidelines, it is considered to be incomplete and lacking clear categories e.g. customization. Some guidelines under the category affordance might better fit under other categories, such as effectiveness or trust. Furthermore, they highlight to keep a consistent detail level across all sections, currently the guidelines in the category effectiveness are more descriptive than others. To provide actionable and usable guidelines, we should consider to frame each guideline in a way that allows a designer to quickly assess whether it has been applied, ideally with a clear yes or no test. Additionally, they suggest including more ”do” and ”don't” examples. For guideline 4 regarding feedback, they suggest adding clarity so that it explicitly states that the system should have elements for users to give feedback and for designers to access it. Referring to the section understanding the challenge, ”how will I know I have understood the challenge referring to model and architecture?”, add what differences it even could be in the examples e.g. classic AI, generative AI etc. Lastly, clearly describe the roles of different actors, including the system, design, and users, to avoid ambiguity.

The guidelines were refined based on the feedback and are presented as the final version in section 6.2 in the result chapter.

## 5.8 Thesis Writing

In this thesis the AI-tool ChatGPT has been used as support for improving sentence structure and looking up synonyms of specific terms to ensure a text with good flow and academic tone. Regarding improved sentence structure, the tool has been used to suggest more concise versions of our composed sentences. The suggestions have been critically evaluated and modified by us to ensure that any information used from the generated suggestion is correct according to the original source. Regarding looking up synonyms the tool has been used by requesting a list of synonyms of a specific term, and if any suitable suggestions the term has been used in the report. Addressing that the thesis work included sensitive and private organizational data, ChatGPT has been used carefully and responsible, ensuring that sensitive data not have been shared.

# 6

# Result

This chapter presents the findings from the research conducted to explore what factors should be considered when seeking to improve AI-driven complex knowledge systems with natural language interfaces in B2B contexts, and how UX design can address these factors. The results reflect insights gathered from a survey, user interviews, user feedback data logs, conversation data logs, and expert evaluations. The chapter begins by presenting an overview of the identified factors, capturing key insights across all data sources, and concludes by presenting the final design guidelines.

## 6.1 Overview of Factors Emerging from Thematic Analysis

To address RQ1 *What factors should be taken into consideration when seeking to improve AI-driven complex knowledge systems with Natural Language Interfaces in B2B contexts?*. The findings from the thematic analysis indicate eight recurring themes. The revealed themes from the thematic analysis were evaluated from an overarching perspective to define factors. Similar themes identified across the different thematic analyses were grouped into one common factor, as they reflected the same underlying concept. In contrast, some themes had a more direct and explicit connection to a single distinct factor. The connection between user feedback comments and specific sessions allowed for deeper analysis of the factors identified from the user feedback log data, revealing patterns that describe underlying causes and challenges influencing the specific factor. The list below presents the factors and associated patterns derived from the results of the questionnaire, interviews, user feedback data, and conversation data. These findings reflect how users perceive and experience the use of a natural language interface in a complex B2B context. Together, these factors aim to represent a diverse set of UX challenges that need to be addressed when improving AI-driven, natural language interfaces in complex B2B settings.

**Collection of Factors and Associated Underlying Patterns**

1. Customization and Flexibility
2. Effectiveness and Efficiency
3. Workflow Considerations
4. Transparency and Trust
   - Traceability Issues
   - Unreasonable Response
   - Outdated Information
   - Hallucinations
5. Language Model Performance
6. Miscommunication
   - Incorrect Scope
   - Incorrect Timeframe
   - Incorrect Level of Detail
   - Abstraction Mismatch
   - Missing Contextual Carryover
7. Level of Detail
   - Response is too generic to match the user's intention.
   - Response is considered to be incomplete or lacking thoroughness.
   - AI provides details that the user considers to be irrelevant.
8. False Affordances
   - Couldn't get the AI to apply the asked timeframe.
   - Unsupported module data access.

## 6.2 Design Guidelines

To address RQ2: *What role can UX design play in addressing factors influencing the improvement of AI-driven complex knowledge systems with natural language interfaces in B2B contexts?* This section presents proposed design guidelines, addressing the eight factors identified to influence the user experience of AI-driven complex knowledge systems with natural language interfaces in B2B contexts. First an introduction on when they should be applied is presented, followed by four categories of guidelines focusing on different goals.

# Design Guidelines for Natural Language Interfaces

These design guidelines aim to support the design of natural language interfaces (NLI) used for information retrieval in B2B contexts. In the context of designing NLIs the term refers not only to the text or voice input mechanisms, but also to the supporting graphical user interface (GUI) elements like buttons, filters, and sliders.

**Who are they for?**

- UX Designers

**When should they be applied?**

- When designing a new NLI
- When improving the design of an existing NLI

**Definition of AI**

Throughout these guidelines, AI is used to refer to both the language model and its underlying architecture.

**Understanding the Challenge**

1. As a UX designer, consider the technical aspects. Explore the scope of the design challenges by understanding the AIs model and architecture capabilities and characteristics.

- What type of AI are you designing for? *E.g. Generative AI, Classic AI*
- What characteristics do the AI have? *E.g. Assertive, Adaptive, Reasoning*
- Do the AI have contextual awareness? *E.g. The ability to remember a users previous questions in a conversation.*
- What level of natural language does the AI understand? *E.g. Can it understand frustration, sarcasm, or slang? How is CAPS LOCK input interpreted?*

2. Use continuous user feedback to adapt to evolving needs.

- Can users leave feedback?
- Is the feedback element placed to be immediately noticeable by the user?
- Can the feedback continuously be accessed by designers?

### Provide User Control and Customization

3. Provide users with the ability to customize search settings. *E.g. Timerange and Data Sources.*

4. Provide users with the ability to adjust the level of detail in responses. *E.g. Generic information or Deep insights*

5. Provide users with the ability to manage and edit specific parts of longer structured outputs. *E.g Modify or delete a specific paragraph.*

### Enhance Trust and Transparency

6. Provide users with the AIs reasoning process in a human readable way. *E.g. Present the steps, logic, and key factors.*

7. Provide users with traceable sources that the response is based on. *E.g. Visible footnotes with links, enabling users to assess the reliability of the provided information.*

8. Provide users with an indication of AIs potential for errors or uncertainty. *E.g. Visible displayed in the GUI or communicated in the response if the AI has limited information.*

### Communicating AI Abilities and Limitations

9. Provide users with an overview of what kinds of expressions the AI can understand and act on.

10. Provide users with appropriate feedback on their inputs. *E.g. Help the user understand how their query was interpreted and how it could be improved.*

11. Provide users with information about if the AI has contextual awareness, and how long context is retained. *E.g. Whether it spans a single session, individual user history, or organizational history.*

---

This result is grounded in real-world needs and challenges, as well as established theoretical frameworks. The theoretical foundation and relevance of the guidelines are further examined in the discussion chapter, where they are connected to existing literature and frameworks.

# 7

# Discussion

This chapter reflects on the findings from the research conducted to explore what factors should be considered when seeking to improve AI-driven complex knowledge systems with natural language interfaces in B2B contexts, and how UX design can address these factors. The chapter begins by discussing the result and describing how the guidelines evolved. Following this a reflection on the used methodologies and execution is presented. Lastly, sections discussing the generalizability of the result, ethical concerns, future work, and contributions are presented.

## 7.1 Reflecting on the Result

This section discusses the identified factors in relation to the derived guidelines. Additionally, connections will be drawn to existing research to highlight alignments and deviations, offering a broader perspective on the results.

In the literature study, several papers and frameworks focused on outlining factors that influence user trust [3], perceived affordances [15]–[17] and how effectiveness can be achieved [26], [30] in AI systems. However, there is a lack of focus on doing so within the complex environment of B2B [7]. The results of this study provided a structured understanding of what factors should be considered when designing AI-driven knowledge systems with NLI in B2B contexts. Through the analysis of user feedback, conversation logs, and survey responses, several recurring factors were identified as common user needs and challenges across Recorded Futures broad user group.

### 7.1.1 Designing for Trust

*Transparency and Trust* was identified as one of the most recurring factors among the conducted analysis. Findings in the literature study shows that trust and transparency have been extensively discussed in existing frameworks, such as HCAI [25], XAI [18], [19] and HCXAI [20], [21]. Our findings reveal that some users experience issues with trusting the AIs responses due to the information in the output from the users perspective being outdated, unreasonable, or possible hallucinations. Examples of users methods for validating the reliability of the output have shown to be through provided references or previous knowledge. XAI builds on the idea of fostering trust through enabling a transparent design, enhancing the users understanding

of how the model works [18]. However, this implies that the user is expected to carry a range of technical skills in order to understand the complex structure of a language model. On that note, Liao and Sundar [3] highlights that users' technical expertise might vary and that developers and designers must understand how users process and perceive information to be able to deliver trustworthy AI-systems.

Another finding from our analysis showed that trust issues arise from not knowing where the provided information was derived from due to traceability issues. This insight highlights a specific user concern as the need to verify where the AI's answers come from and what sources they are based on to be able to validate the AIs accuracy and build trust.

While UX designers cant modify the technical capabilities of the AI, they can build trust through providing a design offering an overview of the reasoning process, highlight potential uncertainties, and include traceable sources to align the users perceptions of the AI with the capabilities. The following guidelines are proposed to support this:

6. Provide users with the AIs reasoning process in a human readable way. *E.g. Present the steps, logic, and key factors.*

7. Provide users with traceable sources that the response is based on. *E.g. Visible footnotes with links, enabling users to assess the reliability of the provided information.*

8. Provide users with an indication of AIs potential for errors or uncertainty. *E.g. Visible displayed in the GUI or communicated in the response if the AI has limited information.*

The first guideline emphasizes the importance of providing the AIs reasoning process to foster trust through transparency. However, Lee argues that transparent models do not necessarily result in a user centered AI solution [21]. Only offering transparency into the reasoning process overlooks the consideration that the users might not have the technical knowledge to interpret the reasoning process, as well as possibly be an inefficient solution. To account for this, HCXAI can be applied by asking questions such as who needs the explanation, why they need it and how they will use it in order to consider the user experience [21]. Therefore the guideline explicitly states that the reasoning process should be presented in a human readable way, aiming to design for a broader group, limiting the technical language. This also aligns with the European Commission's recommendation of considering the involved stakeholders and making the AI decisions understandable for them [24].

The second guideline addresses the users' need to validate the reliability of the information by providing access to the sources used to generate the output, as well as addressing the importance of transparency and human control through traceability mechanisms [24]. While our findings suggest that additional validation options, such as identifying whether a source is the original or indicating the reliability level of sources, could be beneficial it presents challenges. The criteria for assessing the validity of a source can vary significantly between users. Given the variability, we recommend that the design focus on providing the necessary tools to facilitate

*Effective and Efficient* validation, without using a single, predefined measure of reliability. This approach ensures transparency and traceability through letting the user assess according to their specific criteria.

The third guideline was developed to ensure that limitations of the system are communicated, as according to Liao and Sundar trustworthiness is shaped by individual perceptions and triggered by trustworthy cues. This highlights the risk that misleading signals can create a false sense of trust, leading users to believe inaccurate information. This challenge is connected to the first section of the guidelines targeting *Understanding the Challenge*. For designers, this means understanding the specific characteristics of the AI, such as the potential for assertive responses that might present uncertain or inaccurate information in a confident tone. If users are unaware of this, there is a risk for overestimating the reliability of the system, leading to misunderstandings and potential misuse.

### 7.1.2 Miscommunication in Natural Language Interactions

The results of this study indicate that five patterns, presented in section 5.5.2.1, are underlying causes of Miscommunication between users and the AI-driven system. Four of these patterns (*Incorrect Scope, Incorrect Timeframe, Incorrect Level of Detail, Abstraction Mismatch*) are related to mismatches between user input and system output, where the AI fails to interpret the users input as intended, leading to ineffective and inefficient use cases. These findings are in line with the work of Do et al. [30] highlighting abstraction matching as a significant challenge interacting with a NLI. This emphasizes the importance of Do et al. proposal of designing for supporting collaborative refinement between the user and AI. Furthermore, continuous use cases with miscommunications due to abstraction mismatch Do et al. means could contribute to user frustration and affect users' technology acceptance negatively, possibly leading to abandonment of the system. These effects were evident in the analysis, where several use cases involving miscommunication showed clear signs of frustration or clear examples of use cases where users abandoned sessions without achieving their intended outcomes in terms of level of detail, scope, or applying timeframe requests. Furthermore, one of the interviews revealed the need to have greater control regarding modifying the output, specifically related to report creation. To effectively address and reduce miscommunication through design, we suggest the following guidelines:

> **10.** Provide users with appropriate feedback on their inputs. *E.g. Help the user understand how their query was interpreted and how it could be improved.*
>
> **3.** Provide users with the ability to customize search settings. *E.g. Timerange and Data Sources*
>
> **4.** Provide users with the ability to adjust the level of detail in responses. *E.g. Generic information or Deep insights*
>
> **5.** Provide users with the ability to manage and edit specific parts of longer structured outputs. *E.g. Modify or delete a specific paragraph.*

The first guideline supports collaborative refinement between the user and AI, aiming to reduce miscommunication and support effective use cases through an approach that still offers a low barrier to use the system, but provide guidance on what should be adjusted to enhance the output. This means that the user can give an input including unsupported expressions, but still achieve their goal through the guidance.

The second guideline covers the users need to *Customize* timerange and data source settings. For Recorded Futures users, these parameters are critical, as they significantly influence the relevance and reliability of the output. However, it is essential to recognize that, while these aspects are crucial for this user group, they may be less relevant in NLIs designed for other user groups with different goals. Therefore the specific parameters are listed as examples and not mentioned in the actual guideline, aiming to highlight the need to provide customization options tailored to the specific requirements of the target user group, ensuring a more effective user experience.

The third guideline is developed to support users in effectively achieving their intended level of detail in the output without miscommunication. *Level of detail* as a factor was not only evident in the miscommunication data but also consistently highlighted across the survey responses, user feedback comments, and conversation data. The analysis revealed that users often seek varying levels of detail in their outputs. For some use cases, a concise summary is the intended goal, while others require more detailed, in depth information. This highlights the need for flexible customization options to accommodate diverse user expectations and prevent potential miscommunication.

The fourth guideline is developed to enhance the *Flexibility* of outputs by allowing users to modify specific parts or paragraphs without regenerating the entire output. This reduces the risk of miscommunication and ineffective use cases, as it minimizes the chances of the AI misinterpreting the intended scope of the change and ensures that well generated sections remain unchanged. Furthermore, it aligns with the HCAI framework, which emphasizes the importance of designing systems offering both high level human control and high level of automation, aiming to create reliable, safe, and trustworthy systems [25].

The fifth pattern (*Missing Contextual Carryover*) identified as a cause of miscommunication concerned the users incorrect perception of the system having contextual awareness. Possible reasons for this perception could be founded in the way the interface is designed, signaling an ongoing conversation, or expectations based on previous interactions with NLIs. In several cases, the users input indicated an assumption that the AI retained context of previous exchanges, resulting in abstraction mismatches when the AI instead interpreted each input as an independent request. This underscores the importance of designing interfaces that clearly communicate the AIs capabilities and limitations, for *Effective and Efficient* system use. To address this challenge, the following guidelines are recommended:

> **11.** Provide users with information about if the AI has contextual awareness, and how long context is retained. *E.g. Whether it spans a single session, individual user history, or organizational history.*

**9.** Provide users with an overview of what kinds of expressions the AI can understand and act on.

The first guideline focuses on providing users with the necessary conditions for effective and efficient system use. Given that the capabilities and characteristics of AIs can vary significantly depending on their underlying architecture, it is essential for the design to align with the capabilities of the AI. For example, offering contextual awareness may be limited in cases where customer data privacy is a priority. In cases where contextual awareness is not supported we agree with the proposed requirement by the European Commission [24] that it is critical to communicate the AIs capabilities and limitations to the users. This transparency supports setting accurate user expectations and reducing the risk of miscommunication and frustration that occur when users perceive something is supported when it is not. Additionally it is of importance to communicate what specific expressions the AI can understand and act on, covered in the second guideline.

### 7.1.3 The Scope of the Guidelines

As stated in the scope and limitations section, the users experience of an AI system is closely intertwined with the underlying AI model and architecture. This presented a challenge during the process of deriving design guidelines as the user experience is influenced by a combination of factors including the language model, system architecture and design decisions. To ensure that the guidelines provide a practical value for the UX designers, they were initially developed and structured according to the key focus areas in this research. This approach was chosen because designers have the potential to make a meaningful impact within these areas according to finding in the literature study.

The evaluation of the initial set of guidelines revealed that they were written in an inconsistent level of detail, suggesting that a rewritten version would benefit designers more as they would not have to rely on their interpretation. Taking this into account, the guidelines were refined before further evaluation, aiming to achieve a consistent level of detail.

The second iteration of guidelines was presented to the experts panel 1 and 2 where each panel separately got to evaluate them. Expert panel 3 was later on also introduced to both the second iteration of guidelines and a third version refined according to the feedback from panel one and two.

Feedback evaluating the second iteration of guidelines revolved a lot around the categorization of the guidelines, effectiveness was perceived as a problematic term, due to it being highly influenced by the individual's perception. Expert panel one suggested that the categories affordance and trust cover the effectiveness aspect, meaning if you do affordance and trust right, you will achieve effectiveness. Expert panel three presented a contradicting view meaning that affordance and trust are not the only aspects needed to achieve effectiveness. This feedback led us to think about the additional challenge regarding the category trust, since trust also is shaped by the individual's perception, it can't be guaranteed but rather enabled through credible

cues [21]. The headlines *Affordance, Trust* and *Effectiveness* were removed to avoid being interpreted as a guarantee but rather a goal. The change in the headlines also aimed to clearly connect the context of the guidelines to a clearer headline, as the third expert panel argued that several of the guidelines under affordance were not actually perceived as an affordance by them. Discussion within the panel revealed that affordance also could be a problematic word due to the different perceptions of what it is.

Panel one expressed the importance for the designer to understand the models capabilities and architecture and account for it in the design. Therefore, the guidelines underwent a refinement after the second evaluation where the section *Understanding the challenge* was introduced to foster an understanding of the model as well as the users. In the third and final evaluation, these changes were appreciated but still carried an uncertainty for the experts on how to interpret them correctly. Drawing upon input from the experts, the *Understanding the challenge* section was refined into two focus areas of understanding the technical aspects and continuously addressing evolving user needs. Both areas were supported with questions that intend to guide the designer in asking questions that will form a better understanding of the design challenge. This thereby supports the designer in translating the guidelines to their AI systems capabilities and the complexity of user needs within a B2B context. The final guidelines advocate for a collaboration between the designer and AI development team as we can not guide the UX designer in the technical aspect of the models architecture but enable guidance in outlining the scope of the challenge before making design choices that provides clear system affordances, trustworthy cues and effective workflows.

## 7.2 Reflection on Methodologies and Process

This section is dedicated to reflect on the overall process and execution of the thesis by explaining and examining the chosen approach and the methods applied throughout the study.

This study included multiple collections of data provided by the company, which consisted of user feedback data logs, conversation data logs, and data from a survey and two interviews. The volume and variety of the data provided a rich foundation for the analysis, capturing the complexity of a diverse user group operating in a real world B2B environment. However, the breadth and depth of the dataset also came with challenges, particularly in terms of time and analytical scope. Due to the time constraints and limited direct communication with Recorded Futures customers, less time was put into more traditional user research methods, such as additional interviews or observational studies. However, the possibility to attend two interviews that were already scheduled by Recorded Future made it possible for us to understand the findings from the questionnaire on a detailed level, revealing specific user needs. Based on the insights gained from the overall data collection, we chose not to adopt Alabbas and Alomars [26] quantitative approach to measuring system effectiveness and efficiency. As the primary focus was to understand which factors are essential when designing for an NLI, we prioritized understanding users

perceptions of the AI. Therefore we questioned whether quantitative metrics of error rate and response time would enhance our understanding of user needs beyond what was captured in the log data. Given more time, it would have provided us with a clearer view of the AI-systems capacity, but we argue that the meaningful insight for this research lies in the complexity of diverse users, leading us to focus on qualitative data.

To further address the choice of methods, additional interviews conducted by us might have offered more direct control over the data collection. However, they would likely have struggled to capture the same diversity and scale of diverse user perspectives as the available data logs did. Instead a great focus was put into the analysis of the data logs. By prioritizing a comprehensive thematic analysis of over 192 feedback comments, 98 conversation sessions, and 39 user survey responses, a foundational understanding of users challenges and needs was established. The available user feedback data set spanned over one year, during which the NLI features evolved in parallel with new software releases. To make sure that the comments were providing valuable insight on current user challenges, the analysis was limited to comments submitted within the last 6 months. While this narrowed scope excluded some potentially valuable user feedback, it allowed for the identification of recurring patterns across diverse user interactions and important factors for a desired NLI without having to account for large differences in the older versions of the AI.

Given more time, it would have been valuable to also examine how the user feedback evolves simultaneously with the development of AI and design features. Such insights could have offered a better understanding of how users' attitudes might change towards the AI and potentially reveal shifts in perceived level of affordance, trust and efficiency dependent on design changes. Additionally, the logged user feedback only captured negative qualitative feedback and the related sessions in the conversation data log was focused on to reveal underlying patterns and causes of the negative experience, therefore this study offers limited insights into factors and guidelines that could be derived from positive user experiences.

While the initial plan accounted for three weeks of analysing the data, the analysis phase required two more weeks due to the extensive data set and delays in gaining access to applications needed to conduct the analysis. For security reasons, use of alternative cloud based tools were not permitted since the sensitive customer data required to be stored and analysed locally on our computers. As a result, the analysis phase overlapped the evaluation phase, resulting in less time for the experts to evaluate the guidelines. Therefore, additional adjustments were made to the evaluation phase to ensure the remaining project schedule could still be completed within the available time.

To validate the guidelines derived from the data collection and literature study, the expert evaluations were conducted. This was carried out using a modified version of the Delphi method, which enabled three additional iterations during the guideline development process. Given the methods flexibility and its applicability at various stages of research [46], it served as a valuable tool for validation in our study. By modifying the segment in which expert advice was collected, we ensured that a

sufficient breadth of expert opinions and feedback was gathered within the time constraints of the thesis. Although our approach differed from the traditional Delphi model, the repeated feedback rounds still allowed us to incorporate a wide range of expert perspectives. In the end, our adapted method stayed aligned with the main goal of the Delphi technique, which is to improve knowledge through structured input from experts[46], while still fitting the practical limitations of our study.

## 7.3 Reflecting on Validity and Generalizability

The data collection used in this research included input from the company' users, which helped capture a range of perspectives and experiences within our scope of the B2B context. While this increases the credibility of data collected from a diverse user group it also limits our findings to interaction with Recorded Futures AI. The insights derived from the data logs provided the research with information of how users interact with the system in real life settings, without the impact of knowing they were being observed. Since all the datasets were automatically anonymized, the risk for selection biases could have been reduced since the result did not survey a certain group of users within the dataset. However, there is no guarantee, as we are unable to verify whether the feedback sessions originated from different users. Although there were differences in writing styles, questions asked and time stamps did give an indication that the data was generated from different users, organizations, and time zones.

During the process of refining the initial set of guidelines, it became clear that some guidelines were perceived as highly product specific. To address this, the final guidelines were formulated in a more general manner, with examples often drawn from Recorded Futures NLI to increase their generalizability. Additionally, we chose to explicitly focus on guidelines supporting NLIs used for information retrieval, emphasizing that while these guidelines are broadly applicable within this context, they may not be fully aligned with other contexts using NLI solutions.

## 7.4 Ethical concerns

Ethical concerns are a critical aspect of research, particularly when involving user data. This study seeks to improve the user experience of AI-driven complex knowledge systems for diverse professional users. Given the sensitivity of professional data, including the data from participants and related organizations it is essential to consider related ethical issues. This section outlines the specific ethical concerns considered in this study and how they were addressed.

Given the potential for sensitive organizational and personal data, an NDA was employed between us as authors of this thesis and Recorded Future to prevent unauthorized data sharing. To ensure confidentiality for the participants, all data throughout this thesis has been anonymized. Quotations and example conversations from users were not included in the report, instead the findings were presented through the themes emerging during the analysis, ensuring anonymity. According to the log

data method, we ensured that the data files were anonymized. Furthermore, the participants of the evaluation were informed that their responses will be anonymized in the report. Regarding data security, all data was stored on Recorded Futures servers with restricted access to authorized people only.

## 7.5 Future work

This research has aimed to address which factors should be considered when seeking to improve an AI-driven complex knowledge system with NLI in a B2B context and delivered a set of design guidelines that aim to address these factors within UX design. The findings in the study are primarily based on data reflecting how Recorded Futures users interact and interpret their NLI, supported by related literature and expert advice from employees within the company. To continue the research and explore more potential influencing factors, future work should focus on addressing if any additional aspects could be captured outside this study. Since AI systems can have different underlying models, architectures and level of maturity that might pose additional factors, a continuous research within different types of AIs would be valuable to expand the set of guidelines and reach more potential stakeholders.

Future studies could also explore how these guidelines can be expanded beyond information retrieval AI systems, helping to close the gap in actionable design guidance for AI tools used in B2B contexts.

## 7.6 Contributions

This research contributes to addressing the existing gap of concrete UX design guidelines for AI-driven systems for professional users within the B2B context. By exploring existing user needs and challenges from real world scenarios, we managed to translate them into concrete factors that were later on derived into guidelines. The guidelines are intended to support UX designers in creating natural language interfaces (NLI) that focus on information retrieval. By aiming to guide the designer in enhancing system affordance, perceived trust and accomplished effectiveness the design guidelines key focus is to understand the challenge and incorporate design elements that enhances the transparency, traceability, customization and reliability of the system.

By involving 15 experts to evaluate the guidelines from a professional standpoint as UX designers and AI developers, we were able to verify that the guidelines are not only interpretable but also add value to design implications. This expert validation strengthens the credibility and practical relevance of our contributions and demonstrates them as a useful resource when designing for an AI-driven system in a B2B context.

# 8

# Conclusion

This thesis aimed to explore what factors UX designers should consider, and how to address these factors when designing natural language interfaces for information retrieval used by diverse professional users, in B2B contexts. The purpose was to address the research gap regarding existing UX design guidelines, which often overlook the specific needs of professional users in work related contexts. Combining insights from users, through logged data, survey responses, and interviews, with insights from evaluations with experts within UX design and AI development, this study provides a structured set of guidelines for designers working with NLIs.

The research identified eight factors, presented in section 6.1, that influence the user experience of NLIs in professional settings. A crucial insight from the evaluations is the critical role of the designer regarding understanding the technical aspects of AI. The findings underscore that the different characteristics and underlying architecture of AIs significantly impact the user experience, making it essential for designers to consider these technical dimensions when designing NLIs.

To address these challenges, the study proposed eleven design guidelines in section 6.2 for enhancing user control and customization, fostering trust and transparency, and effectively communicating the capabilities and limitations of the AI. These guidelines address the critical challenges of creating user trust and minimizing miscommunication, which emerged as primary challenges in the study.

These findings contribute to both academic knowledge and practical design strategies for NLIs in professional contexts. The identified factors and proposed guidelines can serve as a foundation for further research within UX design for AI-driven systems in professional environments. Future research could expand on this study by validating the guidelines across different B2B sectors and explore possible additional factors. Given the diversity in AI systems, with varying underlying architecture and characteristics, continuous research within different types of AIs would be valuable to expand the set of guidelines.

In conclusion, this thesis provides a step towards bridging the gap in UX design research for AI-driven systems with NLIs for professional contexts, offering practical insights that can enhance the overall user experience through designing for trust, and effective information retrieval that facilitates user friendly interactions.

# Bibliography


[1] O. Badmus, S. A. Rajput, J. B. Arogundade, and M. Williams, "Ai-driven business analytics and decision making," *World Journal of Advanced Research and Reviews*, vol. 24, no. 1, pp. 616–633, 2024. DOI: 10.30574/wjarr.2024.24.1.3093. [Online]. Available: https://doi.org/10.30574/wjarr.2024.24.1.3093.

[2] S. Meshram, T. More, N. Naik, S. Kharche, and M. VR, "Conversational ai: Chatbots," in *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, 2021, pp. 1–6. DOI: 10.1109/CONIT51480.2021.9498508.

[3] Q. V. Liao and S. S. Sundar, "Designing for responsible trust in ai systems: A communication perspective," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 1257–1271. DOI: 10.1145/3531146.3533182. [Online]. Available: https://doi.org/10.1145/3531146.3533182.

[4] S. Amershi, D. Weld, M. Vorvoreanu, *et al.*, "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, 2019, pp. 1–13. DOI: 10.1145/3290605.3300233. [Online]. Available: https://doi.org/10.1145/3290605.3300233.

[5] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "Ai-driven cybersecurity: An overview, security intelligence modeling and research directions," *SN Computer Science*, vol. 2, no. 3, p. 173, 2021. DOI: 10.1007/s42979-021-00557-0. [Online]. Available: https://doi.org/10.1007/s42979-021-00557-0.

[6] RecordedFuture, *Recorded future: Threat intelligence*, https://www.recordedfuture.com/, Accessed: 2025-04-21.

[7] P. S. Caglar, V. Roto, and T. Vainio, "User experience research in the work context: Maps, gaps and agenda," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–28, Mar. 2022. DOI: 10.1145/3512979.

[8] International Organization for Standardization, *ISO 9241-210:2019 Ergonomics of Human-System Interaction Part 210: Human-Centered Design for Interactive Systems*, Geneva, Switzerland, 2019. [Online]. Available: https://www.iso.org/standard/77520.html.

[9] J. Girard and J. Girard, "Defining knowledge management: Toward an applied compendium," *Online Journal of Applied Knowledge Management*, vol. 3, no. 1, pp. 1–20, 2015.

[10] I. Androutsopoulos and M. Aretoulaki, "629 natural language interaction," in *The Oxford Handbook of Computational Linguistics*, Oxford University Press, Jan. 2012, ISBN: 9780199276349. DOI: 10.1093/oxfordhb/9780199276349.013.0035. eprint: https://academic.oup.com/book/0/chapter/293288295/chapter-ag-pdf/44512867/book_34563_section_293288295.ag.pdf. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780199276349.013.0035.

[11] A. de Barcelos Silva, M. M. Gomes, C. A. da Costa, *et al.*, "Intelligent personal assistants: A systematic literature review," *Expert Systems with Applications*, vol. 147, p. 113 193, 2020, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113193. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420300191.

[12] H. Voigt, Ö. Alacam, M. Meuschke, K. Lawonn, and S. ZarrieSS, "The why and the how: A survey on natural language interaction in visualization," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 348–374. DOI: 10.18653/v1/2022.naacl-main.27. [Online]. Available: https://aclanthology.org/2022.naacl-main.27.

[13] W. H. DeLone and E. R. McLean, "Information systems success: The quest for the dependent variable," *Journal of Management Information Systems*, vol. 3, no. 1, pp. 60–95, 1992. DOI: 10.1287/isre.3.1.60.

[14] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, Sep. 2003. DOI: 10.2307/30036540.

[15] G. Pozzi, F. Pigni, and C. Vitari, "Affordance theory in the is discipline: A review and synthesis of the literature," in *Proceedings of the Twentieth Americas Conference on Information Systems (AMCIS)*, Savannah, USA: Association for Information Systems, 2014.

[16] W. W. Gaver, "Technology affordances," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1991, pp. 79–84. DOI: 10.1145/108844.108856.

[17] D. A. Norman, "Affordance, conventions, and design," *Interactions*, vol. 6, no. 3, pp. 38–42, May 1999. DOI: 10.1145/301153.301168.

[18] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[19] S. Ali, T. Abuhmed, S. El-Sappagh, *et al.*, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101 805, 2023. DOI: 10.1016/j.inffus.2023.101805. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[20] U. Ehsan, P. Wintersberger, Q. V. Liao, *et al.*, "Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI

EA '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–7. DOI: 10.1145/3491101.3503727. [Online]. Available: https://doi.org/10.1145/3491101.3503727.

[21] S. Lee, "Toward hcxai, beyond xai: Along with the case of referring expression comprehension under the personal context," in *HCI International 2023 Late Breaking Posters*, ser. Communications in Computer and Information Science, C. Stephanidis, M. Antona, S. Ntoa, and G. Salvendy, Eds., vol. 1958, Cham: Springer, 2024. DOI: 10.1007/978-3-031-49215-0_5. [Online]. Available: https://doi.org/10.1007/978-3-031-49215-0_5.

[22] J. Auernhammer, "Human-centered ai: The role of human-centered design research in the development of ai," in *Synergy - DRS International Conference 2020*, S. Boess, M. Cheung, and R. Cain, Eds., Design Research Society, online, Aug. 2020, pp. 1315–1333. DOI: 10.21606/drs.2020.282. [Online]. Available: https://doi.org/10.21606/drs.2020.282.

[23] A. A. Khan, S. Badshah, P. Liang, *et al.*, "Ethics of ai: A systematic literature review of principles and challenges," in *Proceedings of the 21st International Conference on Software and Systems Process (ICSSP 2022)*, New York, NY, USA: Association for Computing Machinery, 2022, pp. 91–100. DOI: 10.1145/3530019.3531329. [Online]. Available: https://doi.org/10.1145/3530019.3531329.

[24] European Commission, High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI*. Publications Office of the European Union, 2019, Catalogue number: KK-02-19-841-EN-N, Accessed: 2025-05-15, ISBN: 978-92-76-11998-2. DOI: 10.2759/346720. [Online]. Available: https://data.europa.eu/doi/10.2759/346720.

[25] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022, Accessed: 2 March 2025. [Online]. Available: https://research.ebsco.com/linkprocessor/plink?id=912334a0-a94d-3d11-83ac-ce16e6752dcf.

[26] A. Alabbas and K. Alomar, "A weighted composite metric for evaluating user experience in educational chatbots: Balancing usability, engagement, and effectiveness," *Future Internet*, vol. 17, no. 2, p. 64, 2025. DOI: 10.3390/fi17020064. [Online]. Available: https://www.mdpi.com/1999-5903/17/2/64.

[27] International Organization for Standardization, *ISO 9241-11:2018 - Ergonomics of human-system interaction Part 11: Usability: Definitions and concepts*, 2018. [Online]. Available: https://www.iso.org/standard/63500.html.

[28] D. Georgakopoulos and M. Hornick, "An overview of workflow management: From process modeling to workflow automation infrastructure," *Distributed and Parallel Databases*, vol. 3, pp. 119–153, 1995. DOI: https://doi.org/10.1007/BF01277643.

[29] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt, and P. W. Woniak, "A survey on measuring cognitive workload in human-computer interaction," *ACM Computing Surveys*, vol. 55, no. 13s, 2023. DOI: DOI:10.1145/3582272.

[30] H. J. Do, M. Brachman, C. Dugan, *et al.*, "Grounding with structure: Exploring design variations of grounded human-ai collaboration in a natural language

interface," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, Article 363, 2024. DOI: 10.1145/3686902.

[31] Y. Wadsworth, *Do It Yourself Social Research*, 3rd. London: Routledge, 2011, p. 224, ISBN: 9781003115373. DOI: 10.4324/9781003115373.

[32] J. Rowley and F. Slack, "Conducting a literature review," *Management Research News*, vol. 27, no. 6, pp. 31–39, 2004. DOI: 10.1108/01409170410784185.

[33] R. Patel and B. Davidson, *Forskningsmetodikens grunder: att planera, genomföra och rapportera en undersökning*, 5th ed. Lund: Studentlitteratur, 2019, ISBN: 9789144126050.

[34] J. Bell, *Introduktion till forskningsmetodik*, 5th ed. Lund, Sweden: Studentlitteratur, 2015.

[35] J. Green and N. Thorogood, "Chapter 6: Observational methods," in *Qualitative Research in Health Care*, C. Pope and N. Mays, Eds., Wiley-Blackwell, 2018, pp. 85–104. DOI: 10.1002/9781119410867.ch6. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119410867.ch6.

[36] V. Clarke and V. Braun, "Thematic analysis," *The Journal of Positive Psychology*, vol. 12, no. 3, pp. 297–298, 2017. DOI: 10.1080/17439760.2016.1262613.

[37] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. DOI: 10.1191/1478088706qp063oa.

[38] S. T. Dumais, E. Domeniconi, and D. M. Russell, "Understanding user behavior through log data and analysis," in *Measuring and Modeling the Web*, M. Hagen, J. Hauffa, and B. Stein, Eds., Springer, 2014, pp. 323–341. DOI: 10.1007/978-1-4939-0378-8_14. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4939-0378-8_14.

[39] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it," *Library Information Science Research*, vol. 28, no. 3, pp. 407–432, 2006, ISSN: 0740-8188. DOI: https://doi.org/10.1016/j.lisr.2006.06.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0740818806000673.

[40] T. Miaskiewicz, T. Sumner, and K. A. Kozar, "A latent semantic analysis methodology for the identification and creation of personas," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2008, pp. 1501–1510. DOI: 10.1145/1357054.1357290. [Online]. Available: https://dl.acm.org/doi/10.1145/1357054.1357290.

[41] J. Pruitt and J. Grudin, "Personas: Practice and theory," in *Proceedings of the 2003 Conference on Designing for User Experiences*, New York, NY, USA: Association for Computing Machinery, 2003, pp. 1–15. DOI: 10.1145/997078.997089. [Online]. Available: https://dl.acm.org/doi/10.1145/997078.997089.

[42] S. I. Rasca, K. Markvica, and B. Biesinger, "Persona design methodology for work-commute travel behaviour using latent class cluster analysis," *Multimodal Transportation*, vol. 2, no. 4, p. 100 095, 2023, ISSN: 2772-5863. DOI: 10.1016/

j.multra.2023.100095. [Online]. Available: https://doi.org/10.1016/j.multra.2023.100095.

[43] J. Salminen, K. Guan, S.-g. Jung, S. A. Chowdhury, and B. J. Jansen, "A literature review of quantitative persona creation," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2020. DOI: 10.1145/3313831.3376502. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376502.

[44] T. Howard, "Journey mapping: A brief overview," *Communication Design Quarterly*, vol. 2, no. 3, pp. 10–13, May 2014. DOI: 10.1145/2644448.2644451.

[45] S. Walter, *User Journey Mapping*. SitePoint Pty, Limited, 2022, ISBN: 978-1-925836-49-3.

[46] C. Okoli and S. D. Pawlowski, "The delphi method as a research tool: An example, design considerations and applications," *Information & Management*, vol. 42, no. 1, pp. 15–29, 2004. DOI: 10.1016/j.im.2003.11.002. [Online]. Available: https://www.researchgate.net/publication/220652672_The_Delphi_method_as_a_research_tool_An_example_design_considerations_and_applications.

[47] D. Khodyakov, S. Hempel, L. Rubenstein, *et al.*, "Conducting online expert panels: A feasibility and experimental replicability study," *BMC Medical Research Methodology*, vol. 11, no. 1, p. 174, 2011. DOI: 10.1186/1471-2288-11-174. [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-11-174.