WHAT CONSTITUTES CONVERSATIONAL AI CHATBOT SUCCESS?

– an investigation into finding the KPIs to measure overall performance

LAB University of Applied Sciences,
Institute of Design and Fine Arts, Lahti (LAB)
Bachelor of Culture and Arts, Bachelor’s Degree Programme in Experience and Service Design
2023
Anna Emilia Hedvall

LAB University of Applied Sciences

ABSTRACT

Author	Publication type	Completion year	Number of pages
Anna Emilia Hedvall	Thesis, UAS	2023	47

Title of thesis

What constitutes conversational AI chatbot success? – an investigation into finding the KPIs to measure overall performance

Degree, Field of Study

Bachelor of Culture and Arts, Bachelor's Degree Programme in Experience and Sesrivce Design

Organization of the client

Front AI Oy

Abstract

As investment into Conversational AI chatbots (CAI) surges in the aftermath of Large Language Models LLM becoming a worldwide phenomenon, research and practice are still looking for a unified way to evaluate the business performance of commercial CAI.

This thesis contributes to the development of a standard set of objectives and KPIs measuring the success of commercial CAI, for a Finnish growth firm (Front AI Oy) to support and develop its business processes and client-facing activities.

To identify the optimal set of KPIs for evaluating CAI success, a review of the literature along with in-depth interviews with Front AI clients is conducted. In the development of a prototype of the standard, the thesis takes a design thinking and service design approach to produce a prototype. An investigation into the end user is conducted using service design tools such as contextual interviews, empathy map canvas, and persona. Customer journey mapping and user testing are used to ensure a user-centric result.

The results present a standard set of objectives and KPIs tailored to Front AI's context. The final prototype serves as a visualized guide for Front AI employees in the process of identifying enterprise-level objectives and suitable KPIs for any CAI.

The thesis has both practical and theoretical implications, presenting the field with a set of standard KPIs for commercial Conversational AI chatbots that are yet to emerge.

Keywords

Conversational AI, commercial chatbots, CAI success, performance, objectives, KPIs, service design, design thinking, user-centric, prototype

LAB University of Applied Sciences

TIIVISTELMÄ

Tekijä	Julkaisun laji	Valmistumisaika	Sivumäärä
Anna Emilia Hedvall	Opinnäytetyö, AMK	2023	47

Työn nimi

Keskustelevan tekoälyn arvo – tutkimus yleistä suorituskykyä mittaaviin tunnsulukuihin

Tutkinto ja koulutusala

Muotoilija (AMK), Kokemus- ja palvelumuotoilun koulutus

Toimeksiantajaorganisaatio

Front AI Oy

Tiivistelmä

ChatGPT:n ja laajojen kielimallien nousu aiheutti syksyllä 2022 maailmanlaajuisen murroksen. Keskustelevaan tekoälyyn pohjautuvien chattibottien investointien lisääntyessä, tutkijat ja käytännön tekijät etsivät entistä enemmän tapaa, jolla kaupallisten virtuaalis-ten assistenttien arvoa ja vaikutusta liiketoimintaan voisi mitata.

Tämä opinnäytetyö tutkii kaupallisen virtuaalisen assistentin menestystä mittaavan standardin kehittämistä. Suomalaiselle kasvuyritykselle (Front AI) kehitetyn suorituskykymittariston tavoitteena on tukea yrityksen liiketoimintaprosesseja ja asiakasläh-töistä toimintaa.

Optimaalisen tunnuslukumittariston kehittämisessä hyödynnettiin kirjallisuuskatsausta ja syväluotaavia asiakashaastatteluita. Prototyypin suunnittelussa käytettiin muotoiluajattelua ja asiakaslähtöisiä palvelumuotoilun menetelmiä. Prototyypin tuotan-nossa sovellettiin suunnittelun työkaluja, kuten haastatteluja, em-patiakarttaa, persoonatyöka-lua, ja palvelupolkua. Käyttäjäkeskeis-en tuloksen varmistamiseksi prototyyppi testattiin loppukäyttäjällä.

Tulokset esittelevät joukon tavoitteita, tunnuslukuja ja mittareita, jotka on räätälöity Front AI:n kontekstiin sopiviksi. Lopullinen prototyyppi toimii visualisena oppaana Front AI -työntekijöille, heidän määritellessään yritystason tavoitteita ja sopivia mitta-reita virtuaalisille assistenteille.

Opinnäytetyöllä on sekä käytännöllisiä että teoreettisia vaikutuksia. Työ esittelee alalle joukon virtuaalisten assistenttien kaupallista arvoa mittaavia tunnuslukuja, joita ei juurikaan tutkimuksissa ja käytännössä olla vielä kiteytetty.

Avainsanat

Keskusteleva tekoäly, kaupalliset chattibotit, CAI menestys, suorituskyky, tavoitteet, KPI-mittarit, palvelumuotoilu, muotoiluajattelu, käyttäjäkeskeinen, prototyyppi

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Topic Background

With the release of OpenAI’s ChatGPT, based on the GPT-3-artificial intelligence, in late 2022, the interest in Artificial Intelligence (AI) chatbots and Conversational AI has increased (Kolari and Kallio 2023, 36-37). The release presented an example of a fluently speaking chatbot conversing on the level of a human, grabbing the attention of the public worldwide (Kolari and Kallio 2023, 19). Since the release of ChatGPT, AI chatbots have become increasingly normalized and talked about in popular media – and over the dinner table –, their implications and effects common knowledge.

While this leap seems to have come out of nowhere, AI has existed since the late 1950s – when the term artificial intelligence was first adopted. Since then, the first neural networks have been built, debunked, and developed. While over time AI has found its way to manufacturing, home appliances, and chess games; the revolution of deep learning came about several decades later, in 2012. The 2010s were generally a prosperous decade for AI development with the likes of sophisticated AI bots such as AlexNet and AlphaGo being introduced. While chatbots – often described as digital bots that can communicate with users through text in a fashion much like messaging and chatting – can be argued to have existed long before the release of Google’s BERT natural language processing (NLP) AI model, the move from primitive first-generation chatbots – often referred to as button bots – came only after BERT was made widely available in 2018. (Kolari and Kallio 2023, 19.)

Nowadays, chatbots widely utilize machine learning and NLP AI technologies to understand and answer the end user (Agarwal et al. 2022, 1015). These chatbots are often referred to as Conversational AI chatbots (CAI) due to their ability to communicate with the user in a human-like manner. This distinguishes CAIs from button-based chatbots that present options to the user in the form of buttons, menus, or predetermined rules and do not utilize AI to understand the user and to provide context (Kisling 2022).

This advancement of NLP and machine learning technologies in the last few years has led to highly efficient CAIs and chatbots being introduced to the market, making button chatbots increasingly obsolete (Suhaili et al. 2021). Simultaneously the growing popularity of technological concepts such as low-code solutions (Vilser et al. 2022, iv) and messaging platforms have introduced new opportunities for CAI (Suhaili et al. 2021). Furthermore, the rapid advancement of mobile device use and consequently changing consumer habits have boosted chatbots as an increasingly popular option for firms that wish to adopt AI into their operations.

1.2 Research Problem Background

The pressure to stay relevant and to optimize service operations combined with the surge in interest in Conversational AI and Large Language Models (LLMs) has had a seismic effect on industries and how firms operate. As a result, firms are becoming increasingly aware of their position in relation to this rapid development. Managers and executives suddenly feel as if they are falling behind in the rapid rise of opportunities as the race for Conversational AI dominance ramps up. (McKinsey & Company 2022.)

As customer expectations continue to rise (De Andrade & Tumelero 2022, 242) in the aftermath of the COVID-19 pandemic, 63% of service and customer support executives state; improving their operations as the main goal for the year 2023 (Gartner 2023). With the very public advancement in Conversational AI development, firms are unarguably gravitating toward the opportunities it can provide (De Andrade & Tumelero 2022, 244). McKinsey & Company (2022) reported that since 2018 the adoption of virtual agents or conversational interfaces and NLP has doubled. Consequently, more firms opt to offer chatbots as a customer service (CS) channel among legacy and other channels, as CAIs have proven efficient – with their fast 24/7 responses and familiar user interfaces recognized by end users – in catering to the modern consumer’s needs and desires (Suhaili et al 2021).

However, despite surging interest in CAIs both in research and practice, firms that have adopted CAIs still struggle to sustain them in operation (Lewandowski et al. 2022; Janssen et al. 2021, 2). Since 2018 firms have threaded lightly, conducting pilots and trials into CAI chatbots but as the industry matures it is expected to attract bigger investments – but only if projects prove to yield real benefits that can be measured financially (Kolari and Kallio 2023, 98-99).

While research focusing on evaluation methods for CAIs has recently gained a great deal of attention – especially from specific conceptual, usability, and technical design perspectives (Lewandowski et al. 2023) – very little is known about the overall performance of chatbots. Effects of lacking performance can have dire consequences as demonstrated by the launch of Google's Bard in February of 2023. The release, which included Bard giving the wrong answer to a question, caused Alphabet's market value to plummet by $100 billion equivalent to a drop in the share price of 7% (Sherman 2023.) While the jury is still out on the actual impact of the mistake, it clearly shows the significance of chatbot performance to a business. However, research is still lacking about what the important metrics and Key Performance Indicators (KPIs) are for measuring chatbot performance on all levels.

In the last couple of years, there has been a sharp increase in the number of publications related to chatbots, especially regarding evaluation methodologies (Casas et al. 2020, 280). Research has however mostly focused on creating standards that focus on evaluating chatbot design, conversation quality, and the end user's experience (Lewandowski et al. 2023). The intensified interest in measuring these aspects arises from the emphasis that firms put on customer experience and the occasional survey – for example, this one conducted by chatbots.org (according to Casas et al. 2020, 280) – stating that the majority of users find their interactions with chatbots in a customer service setting less than ineffective (Casas et al. 2020, 280).

As providing a great customer experience becomes vital for firms aiming to develop their service operations, handling customer care and communication (Suhaili et al. 2021), so becomes the need for understanding how chatbots as a tool contribute to that. Firms undoubtedly understand the value and potential of CAIs, as service development is seen among the AI capabilities to have the biggest effects on revenue increases for a firm (McKinsey & Company 2022). This has led to increased interest in the evaluation of CAI chatbot performance (Kolari and Kallio 2023, 98-99).

A fundamental issue persists. Firms do not have AI strategies or data collection infrastructure to support CAI solutions (Kolari and Kallio 2023, 97). While optimization of service operations is reportedly the top use-case of AI adoption alongside the likes of contact-center automation and customer service analytics (McKinsey & Company 2022), firms still seem to lack clear strategies, objectives, and follow-up methods that measure the impact and success of their CAI investments (Kolari and Kallio 2023, 97).

Currently, there are no clear guidelines in research for which KPIs firms should use to measure their CAI's success or impact on business. Some researchers argue that it is not the lack of evaluation methods in general (Russell-Rose 2017) but the multiple different standards emerging and mixing over time that creates ambiguity (Casas et al. 2020, 280-281). According to Lewandowski et al. (2022; 2023) and Meyer von Wolff et al. (2021), the field of research on CAI success, objectives, and KPIs lacks clear theoretical and practice-based knowledge. To date, no leading or well-recognized standard has emerged for CAI and chatbot evaluation (Casas et al. 2020, 281) and the general question remains: What kind of metrics should be applied, and what constitutes success for a Conversational AI chatbot anyway? (Russell-Rose 2017).

1.3 Purpose, Aim, and Objective

The topic of this Bachelor's Thesis is *What constitutes Conversational AI chatbot success? – an investigation into finding the KPIs to measure overall performance*. The thesis consists of two parts: Part 1 and Part 2. Part 1 is an investigation into the academic literature and practical knowledge around Conversational AI chatbot (CAI) success, performance, objectives, and KPIs. Part 2 presents a productive prototype development process in which Front AI Oy serves as the case example.

This thesis was commissioned and produced in partnership with Front AI Oy. Front AI is a growth firm founded in 2019 specializing in the delivery of customer service automation solutions for B2C customers. Front AI Oy's headquarters reside in Helsinki, Finland with one additional point of operation in Sweden. The firm consults organizations regarding customer service channel strategies and chatbot implementations. Front AI's main business is delivering CAI implementation projects alongside providing their customers with services related to the maintenance, development, and improvement of chatbots and customer service solutions.

The commissioner of this research, Front AI Oy, seeks to understand which KPIs firms use to evaluate their CAI's impact on efficiency, customer, and employee experience, and cross- and upselling. As no clear standard or practice-based knowledge exists for evaluating CAI success, Front AI aims to establish an understanding of KPIs that can be used in evaluating CAI success and performance on an enterprise level. The first objective of the thesis is to identify the present evaluation methods and KPIs that measure CAI's success and performance. The identified KPIs should be relevant to Front AI and the context it operates in, supporting the advisory function and facilitating implicit knowledge acquisition.

This thesis is conducted as a part of Front AI's business process and knowledge management effort. The thesis contributes to the development of internal processes around performance reporting and evaluation of CAI. Front AI engages in multiple activities that focus on the maintenance and development of deployed CAIs, offering services to clients that include monitoring and evaluation of certain aspects of their CAI. To perform these activities efficiently, Front AI wishes to further develop internal processes to enhance knowledge sharing cultivating the necessary skills of CAI evaluation. Increasing understanding within Front AI of the aspects that constitute CAI success and performance will help Front AI make more data-driven decisions and align activities to support Front AI's operational goals.

The first research objective of this thesis is: to establish an understanding of what constitutes CAI success in Front AI's context and to map which objectives and KPIs Front AI should use to evaluate CAI chatbots on an enterprise level. Part 1 of the thesis will focus on reaching this objective.

The focus of this objective does not lie on the development of new KPIs but on the discovery and compilation of existing ones based on Front AI's needs. The thesis will examine academic literature and practice to determine the KPIs most suitable for Front AI and the context it operates in. However, defining in itself is not enough to support the internal development of processes and CAI evaluation activities at Front AI. To further knowledge sharing and skill acquisition, the findings should be presented in a way that makes the adoption of the concepts around CAI success, objectives, and KPIs easy. The second objective of the thesis thus encompasses the production of a user-centric prototype solution that will help Front AI develop better business and knowledge management processes around CAI evaluation. Part 2 of the thesis will focus on reaching this objective by utilizing a design thinking and service design approach.

The second objective of the thesis is: to develop a user-centric prototype solution for the purpose of internally sharing information about CAI, success, objectives, and KPIs at Front AI.

This thesis hopes to contribute to the research and practice-based knowledge around commercial Conversational AI chatbot evaluation in addition to providing Front AI with information and methods that will help in the development of business processes, and client-facing activities. In the process of reaching the objectives, the thesis aims to answer the following research questions:

- What are the CAI KPIs mentioned in academic literature?
- What are the objectives and KPIs of Front AI's clients?
- Which KPIs should Front AI be focusing on when evaluating CAI success?
- What is the current understanding of CAI's success, objectives, and KPIs at Front AI?
- How should CAI success, objectives, and KPIs be introduced into Front AI to further their understanding of them?

1.4 A Design Thinking and Service Design Approach

The thesis utilizes a design thinking and service design approach in its research design.

Service design has been defined in a multitude of ways in the past few decades since its popularization in academia and practice as a means to support innovation in service organizations (Clatworthy 2013, 16). Service design largely integrates and builds on the concept of design thinking which is a human-centric approach to problem-solving and innovation (Butler 2018). Service design utilizes for the most part the same methods, tools, and approaches as design thinking but differs in terms of consciously and solely focusing on the design of systems that encompass users and providers of services (Butler 2018). While there is an array of approaches and viewpoints to service design (Clatworthy 2013, 16-17); the core principles of service design have remained unanimous over the years (Clatworthy 2013, 20) often described as human-centered, co-creative, tangible, holistic, and orchestrated (Narges et al. 2018).

Human-centered design encapsulates focusing on the people whom we are designing for while also extending beyond the user to include other stakeholders e.g., businesses and commercial organizations. As this thesis's objective is to develop a prototype solution for the internal use of Front AI, it is important that the final solution is geared towards and takes into consideration the people who will engage with it. Similarly, including clients and other stakeholders – in addition to Front AI key personnel – in the design process, ensures that the solution caters to the needs of Front AI – the target user – that in turn caters to their clients. Including stakeholders in the design and delivery of a service experience constitutes the co-creative core principle of service design. (Narges et al. 2018.)

The tangible core principle of service design describes the process of making an intangible service physical, through digital touchpoints. As Front AI aspires to develop its business and knowledge management processes and to increase understanding of CAI evaluation, the solution needs to display and visualize CAI evaluation adjacent concepts. Digital services are inherently intangible and are defined by their digital touchpoints. Tangible touchpoints can be seen and experienced thus creating value and concrete interactions. Utilizing service design and design thinking in creating the prototype solution ensures that the notion of CAI evaluation is presented and understood by the target users. (Narges et al. 2018.)

The holistic core principle encompasses the end-to-end experience of using the prototype accounting for all steps, not just one single interaction, encounter, or stage in the process. Besides including all touchpoints in the process, the surrounding context must be considered. The holistic core principle ties into the orchestrated core principle which considers the context of where the service delivery will occur, ensuring that the various elements and processes of an organization are optimally aligned. As this thesis plays a role in the grand scheme of Front AI's internal development of its processes and activities around CAI evaluation, the concept should consider the surrounding processes and environments – including the current activities and notions that it aims to further support. The concept should support internal interactions as well as grow the implicit knowledge of CAI evaluation within the firm. (Narges et al. 2018.)

As the thesis aims to bring forward a solution in the form of a prototype for Front AI employees to increase their understanding of CAI success and performance on an enterprise level, the approach should ensure this objective is achieved. As the prototype solution aims to cater to the employees of Front AI, the approach should include methods and tools that are user-centric. As argued above, service design and design thinking include various viewpoints and approaches that enable the thesis to reach its objective.

1.5 Structure of the Thesis

The thesis consists of two parts: Part 1 and Part 2. Front AI serves as the case example in the latter. Part 1 consists of a literature review, Front AI client interviews, and a summary of the findings satisfying the first objective of this thesis. Part 2 introduces the practical part of the thesis and the design process. The chapter will present the stages of the design process; user research, defining the design problem, and the development of a solution. The chapter will also present the final prototype. The final chapter will present the concluding remarks of the thesis, summarizing the research.

1.6 Timeline and Process

The topic of the thesis was first introduced by Front AI in May of 2022. The process of producing the thesis began in September of 2022 after an agreement had been struck with Front AI regarding the objective and purpose of the thesis. The thesis began as a general investigation into CAI business benefits, with client interviews being scheduled for October and November of 2022, to fulfill the commissioner's wish to get preliminary results by the end of the fourth quarter of the same year. Findings from the client interviews were presented in January of 2023 to Front AI.

During the second quarter of 2023, the thesis was repurposed and expanded to focus on the development of Front AI's activities around reporting. By June 2023 the thesis had grown to include the development of a prototype to further the understanding of the concepts around CAI success, objectives, and KPIs collected during the early phases of the thesis process. Interviews with Front AI employees were held from June to August 2023. The development of the prototype took place simultaneously during August 2023. User tests were conducted during August and September, with the documentation occurring simultaneously.

2. PART 1: IDENTIFYING CAI KPIS FOR FRONT AI

This chapter will present the literature review consisting of secondary data on CAI evaluation in academic research. The secondary data was collected using scientific databases such as LAB Primo and search engine platforms like Google Scholar. The selected literature was assessed for validity by focusing on peer-reviewed academic journals and selected based on relevance to the topic and the research question, as recommended by Stickdorn et al. (2018). The literature review scope is structured in line with the research questions and discusses the current theoretical literature landscape of CAI evaluation with an enhanced focus on the evaluation of CAI success and performance on an enterprise level.

In addition to the literature review, this chapter will present primary data from a multiple case study on CAI evaluation, to increase practice-based knowledge. To pinpoint the most central CAI evaluation KPIs from the perspective of Front AI it was deemed essential to include Front AI's clients in the study. Understanding how Front AI's clients with deployed CAI chatbots evaluate the performance of their CAIs contributes to a more nuanced interpretation of the most relevant KPIs for Front AI. The primary data collection took a qualitative case-study approach to discover the objectives and KPIs that Front AI's clients use for evaluating their CAI chatbot's success and performance. In-depth interviews were conducted with representatives of each case company. The collection, analysis, and results of the primary data collection will be presented later in this chapter.

The concluding section of this chapter presents the findings from the literature review and primary data collection. The section discusses and combines the results from each section, synthesizing the findings. Based on the findings, the most commercially viable and central KPIs for Front AI will be identified and presented.

2.1 Literature Review: Conversational AI evaluation

In the last few decades, research has explored methods and philosophies for chatbot evaluation. The first studies focused on dialog systems, which could be everything from IVR and conversational AI (Radziwill & Benton 2017). In the last few years, research on chatbot performance has increased due to the rate at which they have been adopted in contemporary organizations (Lewandowski et al. 2023). More studies have thus started to examine CAIs or other text-based conversational interfaces (Peras 2018, 89) from the point of view of classification, typology, or how different chatbot characteristics and differentiators should be taken into account in the designing and evaluation of such interfaces (Følstad et al. 2019; Peras 2018,).

Three major approaches have been identified in the literature on CAI evaluation. The first approach is based on the discovery that methods used to evaluate CAI generally align with the ISO 9214 concept of usability consisting of three categories: effectiveness, efficiency, and satisfaction (Casas et al. 2020, 281). The second approach presents the same notions but in the form of three main categories: Content evaluation, Functional Evaluation, and User Satisfaction. (Casas et al. 2020, 281) The two approaches overlap heavily, exhibiting largely the same metrics for evaluation but differing in the categorization of them. The third approach encompasses 4-5 perspectives (Jadeja and Varia 2017; Peras 2018) on chatbot evaluation which should be used either individually or together depending on the objectives of the chatbot (Jadeja and Varia 2017 according to Casas et al. 2020, 281)

Studies that have taken the approach of categorizing methods from the point of view of Usability have sorted metrics into three categories: effectiveness, efficiency, and satisfaction (Casas et al. 2020; Radziwill & Benton 2017).

Effectiveness is comprised of two aspects: functionality and humanity (Casas et al. 2020, 281). Effectiveness generally corresponds to a technical point of view where the main point of evaluation is looking at the performance of the underlying algorithms. Evaluating effectiveness means measuring how accurately the chatbot interprets user questions, how it handles unforeseeable questions, what the linguistic accuracy is, and how it measures up to a human-to-human conversation (Casas et al. 2020, 281; Hung et al. 2009). This essentially amounts to how effectively a chatbot can respond to user input (Hung et al. 2009) The assessments of these aspects can be conducted using questionnaires.

Casas et al. (2020, 282) sort the following evaluation methods in literature as ones measuring effectiveness: performance, content evaluation, information retrieval perspective, conversation intelligence, and domain coverage. These methods can be assessed quantitatively (Casas et al. 2020, 282). For example, in literature Content evaluation that focuses on measuring the accuracy of the chatbot's response to a user inquiry, can be assessed using automatic evaluation or expert evaluation (Maroengsit et al. 2019, 115).

Automatic evaluation evaluates the chatbot's responses assessing them on precision and recall which refers to the measurement of how many times the chatbot's responses are relevant to the topic of discussion and the percentage of user questions that the chatbot can correctly match to the corresponding topic (Maroengsit et al. 2019, 115). These assessments can easily be conducted automatically using text summarization methods such as BLEU (bilingual evaluation understudy) and ROUGE (Recall-Oriented Understudy of Gisting Evaluation) that produce quantitative data (Maroengsit et al. 2019, 115).

Expert evaluation focuses on measuring the naturalness and suitability of a response in a conversation. This is most accurately judged by a human. Several frameworks have been created to perform expert evaluation and they mostly consist of qualitative metrics (Vilser et al. 2022 p 15). Human-based evaluation is largely favored for its versatility in evaluating functional aspects as well as aspects related to quality and the determination of whether the user's need has been satisfied (Vilser et al. 2022 p 15). The effectiveness of CAI can thus be assessed using both qualitative and quantitative metrics.

Casas et al. (2020, 282) report a decline in the use of effectiveness and content evaluation methods in recent academic literature. The instances in which methods measuring the effectiveness of CAI have been used are instances of e.g., machine learning, NLP techniques, or when the underlying algorithms have been assessed. It has become increasingly rare to assess CAI only based on effectiveness or technical performance as task-based customer service chatbots interacting with users are becoming more popular because chatbots that help customers should be assessed on more aspects than just technical functionality. (Casas et al. 2020, 282.)

The second category of the Usability-approach focuses on evaluating the efficiency of the CAI. Efficiency looks at whether the chatbot performs the task it is made for successfully and thus fulfills its purpose. Chatbots that focus on achieving a certain goal, referred to as Task-oriented chatbots, tend to be used by organizations that need help with customer service. In these instances, the efficiency of the chatbot is measured based on whether e.g., the chatbot manages to answer the user's inquiry or to fool the user into thinking they are human. (Casas et al. 2020, 282-283.)

Maroengsit et al. (2019) reasons for task-based evaluation as a part of Functional evaluation. Similarly, as in the Usability approach, Functional evaluation consists of evaluating goal-oriented chatbots e.g., chatbots with the purpose of assisting users, on how well they perform said goal. Functional evaluation also encompasses looking at usage statistics to indicate how well the chatbot performs. The usage statistics are largely quantitative, displaying user data on, for example, the number of words or unique utterances (Maroengsit et al. 2019, 117). Casas et al. (2020) classify the following methods mentioned in the literature that measure efficiency as: functionality, humanity, functional evaluation, linguistic perspective, AI perspective, Coherence, Conversation breadth, and depth.

While determining the efficiency of CAI has grown in popularity in the last few years – according to Casas et al. (2020, 283) – the method is rarely used alone as task-based chatbots tend to have goals that are heavily intertwined with user satisfaction, such as helping users with customer service inquiries. In this case, as the CAI's goal is to help with customer service and interacting with customers it subsequently needs to be evaluated on user satisfaction (Casas et al. 2020, 283). Evaluating user satisfaction is especially important in the context of commercial chatbots, where falling short of meeting users' expectations can lead to the failure of the chatbot altogether (Janssen et al. 2021). Satisfying the needs of real users is seen by some researchers as the ultimate goal or objective of any CAI (Vilser et al. 2022). The notion of user satisfaction has thus become central to a lot of CAI evaluation approaches, methodologies, and frameworks (Maroengsit et al. 2019, 116).

User satisfaction focuses on methods that ask the user about their interaction with the chatbot. User satisfaction can be rated on two levels: on an entire chat session basis (session level) or a single message or response basis (turn level). Session-level evaluation relates to how the chatbot is perceived compared to a human while turn-level assessment is based on asking users to evaluate each separate response from the chatbot. Data can be collected using expert evaluation or direct feedback from users. Evaluators can fill in a questionnaire rating the whole session, on specific aspects of their experience or satisfaction or rate their experiences of a single turn using a Likert scale. The factors and aspects making up user satisfaction differ between research papers. (Maroengsit et al. 2019 p. 116.)

For example, in their study on customer satisfaction of digital assistants, Brill et al. (2019) introduce a generic expectations-confirmation model to illustrate the aspects that influence a user's satisfaction with an interaction with a chatbot. The aspects are Expectations, Perceived Performance, Confirmation of Expectations, Perceived trust, and Information Privacy Concerns (Brill et al. 2019). Casas et al. (2020, 282) compile the methods mentioned in the literature that account for User satisfaction as: affect, ethics, behavior, accessibility, user satisfaction, user perspective, Chatbot interface, chatbot personality, Conversational user experience, and engagement. These methods account for aspects of chatbot personality, authenticity, respect, privacy, and trustworthiness (Casas et al. 2020, 281).

While the Usability approach provides a holistic approach to chatbot evaluation, current studies still emphasize the importance of choosing evaluation metrics based on the chatbot's purpose (Jadeja & Varia 2017; Cahn 2017 by Peras 2018; Shawar and Atwell 2007 by Hung et al. 2009). As CAIs can be used in numerous areas and for various purposes the evaluation should be based on the needs of the chatbot users and aligned with the area of application (Peras 2018; Jadeja & Varia 2017). For example, Jadeja and Varia (2017, 2) argue that if the CAI is used for business purposes the existing metrics and KPIs need to be modified to reflect properties like user engagement and retention. The evaluation approach and metrics should thus be adjusted to the purpose of the chatbot (Cahn 2017 by Peras 2018). The introduction of the final Perspectives-approach is born from this notion.

2.1.1 A Perspective Approach to Commercial CAI Success Evaluation

The perspectives-approach introduces the idea that CAI performance can be assessed from a multitude of perspectives. All perspectives value different objectives and outcomes e.g., the information retrieval perspective values accuracy, precision, and recall of CAI responses while the user experience perspective values usability and satisfaction (Peras 2018, 93). The perspectives approach classifies usability and user satisfaction as a part of the user experience perspective taking a different approach than the three category approaches (Jadeja and Varia 2017; Russell-Rose 2017; and Peras 2018). Russell-Rose (2017) presents a 4-perspective framework consisting of 1) user experience perspective 2) information retrieval perspective 3) linguistic perspective 4) technology perspective. Peras (2018, 92) proposes a five-perspective approach based on Russell-Rose's (2017) approach, by adding a fifth perspective: business perspectiv, illustrated in Table 1. Peras (2018) introduces the Chatbot Evaluation Framework (Appendix 1) exhibiting the five perspectives with corresponding metrics and approaches to evaluate the specific categories making up each perspective.

Table 1 The Business perspective in the The Chatbot Evaluation Framework by Peras (2018).

Perspective	**Category**	**Attributes**	**Metrics**	**Approach**
Business perspective	Business value	• efficiency cost • qualitative cost	• number of users • duration of the chatbot conversation • number of the chatbot conversations • number of the agents included in conversation • duration of the conversation with an agent • number of the unsuccessful conversations • number of the unsuitable responses • number of repeated queries	Quantitative

Peras (2018) introduces the business perspective as its own perspective in their chatbot evaluation framework. Previous research has not made this distinguishment. The business perspective specifically looks at the CAI effect on business value helping assess its appropriateness and validity in a commercial context. Peras (2018) introduces metrics for measuring business value. The purpose is to assess the value that a chatbot can bring to a business by looking at the quantitative data on the number of users, duration of the chatbot conversation, number of the chatbot conversations, number of agents included in the conversation, duration of the conversation with an agent, number of unsuccessful conversations, number of unsuitable responses, and number of repeated queries. (Peras 2018.)

The chatbot evaluation framework by Peras (2018) addresses the shortcoming in literature, where most studies have focused on evaluation from one perspective not considering chatbot typologies or multiple perspectives (Peras 2018, 92). The business perspective emphasizes the chatbot as a tool to create business value. As the application of a specific perspective or a mixture depends on the type of CAI that is evaluated, it is expected that commercial chatbots that have been deployed to help in customer service should also be assessed on their contribution to the business in this regard, as chatbots tend to fail if not proven commercially viable (Janssen et al. 2021).

For this reason, the examination of commercial chatbots has received some additional attention in literature with the effort to expand on the evaluation methods around CAI success.

In their paper, Lewandowski et al. (2023) specifically focus on commercial CAIs presenting a set of relevant criteria and a model on how to evaluate the quality of CAI and determine the CAI's success holistically. The authors mention Performance consisting of criteria such as effectiveness and efficiency as a strong predictor of CAI success (Lewandowski et al. 2023). The authors define performance to be directly related to user satisfaction. This, the effective and efficient completion of executed tasks in the conversation, is performance. To measure effectiveness Lewandowski et al. (2023) use task (success) rate (Peras, 2018), task failure rate, and retention and feedback rate as effectiveness metrics and task completion time, average number of turns, and human-handover rate to measure efficiency.

Another study focusing on evaluating the commercial success of CAI is De Andrade and Tumelero's (2022) paper investigating the effects of an AI chatbot on increasing the effectiveness of customer service at a bank. The authors argue that their findings show that the adoption of CAI brings significant gains due to automation standardization and optimization of existing processes within an organization reducing costs and improving operational efficiency (De Andrade & Tumelero 2022, 239).

In their interviews with industry professionals De Andrade and Tumelero (2022) found that respondents highlighted CAI's ability to contribute to agility, availability, accessibility, resoluteness, transshipment, and prediction. No actual metrics were mentioned concerning the evaluation of agility, availability, and accessibility other than that it has a positive effect on service efficiency as the CAI can change time-consuming human calls into fast and efficient interactions, which can be described as the estimated latency time. Resoluteness signifying the percentage of customer service that the CAI application effectively answers without the help of a human is said to reduce queues at call centers. Resoluteness excludes the transshipment of the user which refers to the transferring of customers to contact an agent or channel for assistance. Prediction refers to how data collected from CAI can help the organization predict customers' intentions and will, improving the experience and customer service progressively. (De Andrade & Tumelero 2022, 248.).

Respondents in De Andrade and Tumelero's (2022) study also mentioned assessing the increase in interactions, new customer adoption, and problem-solving using metrics such as the number of interactions over a certain time period and a problem-solving index or resolution rate. The rate of transferred calls to human attendants was also determined using quantitative metrics (De Andrade & Tumelero 2022, 246). The CAI was also said to have decreased wait time and freed up human agents to handle more intricate issues (Osei-Mensah et al. 2023, 95) (De Andrade & Tumelero 2022, 247).

2.1.2 Summary of the Literature Review

The literature has presented several approaches to CAI evaluation. The literature on chatbot evaluation started with focusing on evaluating CAIs based on technical performance and functional aspects. Since a more nuanced approach has been introduced with aspects such as efficiency and user satisfaction gaining more attention. While the literature has been highly focused on creating one framework to evaluate all chatbots with, the notion of modifying approaches, methods, and metrics to fit CAI purpose, objective, and typology has gained ground among researchers. The perspectives approach allows for such adjustment ensuring that the right metrics are being used for evaluating chatbots. The identification of the chatbot's objective has thus been highlighted in the literature.

It is only recently that commercial CAIs have received attention despite the surge in investments in this area. Only a handful of papers have discussed the repercussions of CAI success and focused on creating metrics for evaluating the business value of commercial CAI. The literature emphasizes mixing perspectives to gain a holistic understanding of CAI performance. The chatbot evaluation framework with the business perspective introduced by Peras (2018) provides a starting point for compiling the right metrics to evaluate CAI success and performance on an enterprise level.

2.2 Client Interviews

This section will present results from the qualitative multiple case study and the process of data collection. The section will begin by presenting the research design – explaining the qualitative case study approach and presenting the research questions. Thereafter the sampling methods and selection will be presented. Then the process of data collection beginning with a summary of the methods and procedures, will be accounted for. Finally, the data analysis section will explain the methods used, followed by the results from the data analysis.

2.2.1 The Qualitative Case Study Approach

The multiple case study -approach is a qualitative research strategy (Punch 2013, 138) usually used in exploratory studies where the purpose is to gain insight into a topic of interest, focusing on the 'what' and the 'how' (Saunders et al. 2015, 174). The case study approach is advantageous when clarification or a more profound understanding is needed regarding a particular issue, problem, or phenomenon (Saunders et al. 2015, 175). Combined with qualitative research methods deeper insights into issues, interdependencies, and user problems can be discovered about the prevailing case. While the case study is often criticized for its inability to yield generalizable results – due to its nature of detailing a particular subject under investigation – it is generally understood that multiple cases can be used to make judgments of the findings' typicality (Hamel et al. 1993, 34). As the thesis aims to understand CAI success, by mapping objectives and identifying the essential CAI KPIs that contribute to commercial success the exploratory case-study approach serves as the appropriate approach to reach the aim. The case study allows for a user-centric approach ensuring the discovery of the most relevant KPIs for Front AI.

Another argument for taking this approach is that the findings from the literature review should be validated using primary data (Hamel et al. 1993, 16) that will consider Front AI's perspective and needs. The literature review showed that there is little research detailing CAI evaluation methods and metrics focusing on the business perspective measuring performance and success in commercial settings and on an enterprise level. The literature emphasized the importance of perspectives in the evaluation of CAI. Understanding what the firms strive to achieve with their CAI is important to determine the right approach and choose the right metrics. This objective should thus be determined in addition to the KPIs to assess the appropriateness of the chosen methods. The interviews will thus aim to answer the following research questions:

- What objectives and performance targets have firms set for their CAI?
- How are these measured and evaluated?
- What are the KPIs that firms use to evaluate their CAI on an enterprise level?

The multiple case-study -approach using qualitative research methods allows for bridging the knowledge gap in the literature around CAI metrics for measuring success and business performance and answering the research questions.

2.2.2 Sample Selection

To pinpoint the most central KPIs from the perspective of Front AI it was deemed important to include Front AI customers in the study. Understanding how firms that have deployed CAIs into their operations evaluate the performance of their CAIs will ensure that the relevant KPIs for Front AI are included.

After deliberation together with the commissioner, 3 case firms from among Front AI's clients were identified. The sampled firms were selected based on the following criteria: 1) industry 2) how long the firm has had a CAI in production 3) prior knowledge of the firms engaging in some sort of evaluation and monitoring of their CAI 4) the availability of suitable interviewees. The criteria were established based on the research questions defined for the interviews to ensure that the needed data would be acquired (Stickdorn et al. 2018, 25).

After the case firms had been selected, suitable representatives and industry experts were identified within each case firm and contacted via email for an interview. The recruitment yielded 3 case firms with 1-2 interviewees each. The interviewees were selected based on their proximity to the firm's CAI operations. Interviewees were selected based on whether they had participated in the planning, execution, and evaluation of the CAI. Seniority and role within the CAI project were also considered in the selection process. As the relevant stakeholders of each case firm were familiar with Front AI, interviewees were also selected based on Front AI's assessment of who would be suitable and available to be interviewed. The final sample consisted of CAI managers, project managers, and product owners as well as other stakeholders on an operational level with experience in CAI investment and management. The opportunity to recruit additional interviewees was kept open in case more data was needed. The case firms, their descriptions and interviewees are summarized in Table 2.

Before the interviews, the interviewees received some information about the purpose of the interview, interview themes, and material they could prepare in advance. The interviewees were urged to collect any relevant material related to their CAI evaluation practices they would be allowed to share for research purposes. Interviewees were also asked to fill in a form detailing their consent to be interviewed.

Table 2 An overview of the case firms and interviewees selected for the client interviews.

Case firms (3)	**Case firm description**	**Interviewee(s)**
Case A - Public sector vendor	The company is a public sector service provider.	Specialist, previous project manager of the CAI initiative. Product owner, managing the CAI initiative
Case B - For-profit vendor	The company is a benefit and mobile payment service provider. They offer employee benefits for employees, employers, and merchant customers.	Product owner, experienced in managing the CAI at the firm
Case C - Insurance and banking vendor	The company works in the insurance and banking sector offering services for private and B2B customers.	Innovation manager, insight into the introduction of the CAI at the firm

2.2.3 Data Collection

The qualitative research method selected for the collection of primary data was in-depth interviews. In-depth interviews are used for collecting qualitative data from relevant stakeholders such as customers or external experts to gain a deeper understanding of prevailing expectations, experiences, or processes (Stickdorn et al. 2018, 24). In-depth interviews can be conducted in different manners, ranging from structured to unstructured interviews (Stickdorn et al. 2018, 24). A semi-structured approach was selected to give the interview some form with pre-determined questions but with room to ask follow-up questions or allow the interviewee to expand on a certain answer (Stickdorn et al. 2018, 24). Semi-structured interviews allow for an approach where interviews are started with general and broad questions after which more specific and detailed questions can be asked after building a rapport with the interviewee. This allows for the collection of useful data related to the research questions (Stickdorn et al. 2018, 24).

In-depth interviews are usually individual interviews held either face-to-face or online (Stickdorn et al. 2018, 24). The interviews were conducted one-to-one over Teams-videocalls. This option was selected to provide more scheduling options for the interviewees as well as a seamless opportunity for recording the interviews. It was also assessed that no loss of information would occur as the option would still provide the opportunity to observe the body language of the interviewee, which is mentioned as one of the benefits of the in-depth interview (Stickdorn et al. 2018, 24). Each interview was conducted in the interviewee's primary language to allow for the building of trust and rapport between the interviewee and the interviewer.

To support the semi-structured interview process, an interview guideline was created. The guideline was designed to include appropriate probes and questions related to the research questions. The probes and questions were designed to be open-ended and as non-leading as possible. The guideline also allowed for flexibility where the agenda could be changed, and follow-up questions could be posed in response to the interviewee's answers. (Stickdorn et al. 2018, 24-26; Saunders et al. 2015, 403.).

The interview started with some general questions which the interviewee was allowed to present themselves. The interviewees were then asked more general questions about the case firm after which discussions about the objectives and KPIs of said case firm were discussed. The interview was then ended with some concluding remarks on the possibility of conducting further interviews with either the interviewees themselves or with other stakeholders from within the firm, if necessary.

2.2.4 Data Analysis

The processing and analysis of the interview data started with the transcription of the interview recordings. The interviews were transcribed with enough accuracy to ensure that no crucial data, such as specific expressions and wording would be lost in the process. Interview notes were also used in the data analysis process to supplement the recordings and transcripts to ensure accurate interpretation and to reduce the risk of loss of data.

The data in the transcripts were then dissected and segmented, assessing and grouping similarities and differences in the data. The data was coded highlighting the important themes around CAI objectives, targets, evaluation methods, and key performance indicators. The research questions were used to guide the discovery of codes and concepts that highlighted a certain phenomenon.

- What objectives and performance targets have firms set for their CAI?
- How are these measured and evaluated?
- What are the KPIs that firms use to evaluate their CAI?

The data guided the formation of the codes and illustrative quotes were gathered to describe the different 1st-order concepts. The coded data was then arranged into themes. 2nd-order themes were derived from the concepts to classify the type of concept (commercial objective and CAI objective). The forming of the 2nd-order themes that represent the aggregated dimensions from the 1st-order codes and concepts in the data is a method often used in qualitative data analysis as it maintains rigor and assures that the findings are backed up by data. (Gioia et al. 2013.)

The results of the data analysis were documented and compiled into a data table to showcase the raw data and how the codes and themes are grounded in it (Gioia et al. 2013, 26). This was done to aid the final analysis and interpretation of the findings.

2.2.5 Findings and Results

The interviews had an explorative and broad approach to the topic of CAI objectives and KPIs with the objective of collecting data related to KPIs and standards that the case firms use to measure their CAI's performance and business impact. The interviews also uncovered which targets and objectives firms set for their CAI ventures and how they follow up on achieving those goals. The result of the data analysis is presented below.

The interviews yielded results on the objectives and the reasons why the case firms have opted to implement a CAI. Some reasons and objectives were more common than others, being mentioned by one or more of the case firms' interviewees. The interviewees had a clear understanding of why the CAI solution had been implemented which was determined to reflect the objective of the CAI. The interviews also yielded results on the future objectives that the firms have set for their CAI.

Some of the frequently mentioned objectives for adopting a CAI were: using the chatbot 1) as a strategic customer service (CS) channel (complementary, alongside other CS channels), 2) to increase, promote, and support digital self-service (channels), 3) to diversify the CS offering (provide 24/7 text-based service and cater to customer expectations and experience (CX), and 4) to decrease the number of incoming customer contacts (to all CS channels but especially phone). These were interpreted and grouped to fall under the theme of common objectives. Table 3 on the next page compiles the common objectives.

Other objectives mentioned by some of the interviewees were the notion of the objective to expand to different topic areas in cases where the firm offers its service in other languages and topic areas. Another objective was to enable fast reactions and even proactiveness of the CS function by introducing CAI. As interviewees were mainly alluding to these objectives highlighting them as benefits of the CAI solution, they were not deemed significant enough to fit into the common objectives. They were thus themed as other objectives.

The interview data suggests that firms choose to invest in CAI as a measure to support other customer service channels and to provide a diverse customer service offering. The case firms especially mentioned the 24/7 service option that CAI can provide. All case firms offer digital self-service, which impacted the decision to invest in CAI, as it is seen as something that can complement and support the digital self-service experience, helping the end users who choose to conduct their business using the provided online services.

Table 3 CAI objectives identified in interviews with Front AI clients.

CAI objective	Case A	Case B	Case C
CAI as a strategic CS channel (complementary, alongside other CS channels)w	Yes	Yes	Yes
Increase, promote, and support digital self-service (channels)	Yes	Yes	Yes
Diversify the CS offering (24/7 service, text-based communication, CX, and expectations)	Yes	Yes	Yes
Decrease the number of incoming contacts (to all CS channels but especially phone)	Yes	Yes	Inconclusive

Table 4 CAI performance KPIs and metrics.

KPIs and metrics that measure CAI performance	Case A	Case B	Case C
Conversation count	Yes	Yes	Yes
Resolution rate	Yes	Yes	Yes
Unknown messages	Yes	Yes	N/a
End-user feedback ('thumbs-up' and 'thumbs-down)	Yes	No	Yes
Number of intents	N/a	Yes	N/a
Most popular intents	Yes	Yes	N/a

The results of the interviews revealed some findings on how the case firms measure progress on their set objectives. The most frequently mentioned way to follow up on the progress of the CAI was the collection of data on the number of conversations held with the CAI. The interviewees also mentioned monitoring the following KPIs: resolution rate (the percentage or count of solved conversations) and unknown rate (the percentage or count of questions the CAI was unable to answer or solve). End-user feedback was also collected and monitored in the form of a rate or count of 'thumbs-up' or 'thumbs-down' reactions to CAI messages or whole conversations. Other data points mentioned by the interviewees to be included in their reporting processes were the number of intents (roughly how many questions the CAI can answer i.e., the scope) and the most popular intents (the most frequently asked questions). The mentioned KPIs are ones used to measure the performance of the CAI. The purpose of the KPIs is to provide insight into how the CAI is performing on a level limited to CAI performance alone e.g., is the CAI giving correct answers and helping the user to solve its issue. As none of the KPIs mentioned, directly measure the impact the CAI has on the firm's CS function – by making comparisons to other channels – they were thus themed and grouped as CAI performance KPIs and metrics. Table 4 summarizes the CAI performance KPIs and metrics.

Additionally, the interviews yielded results on KPIs that the case firms use for measuring the impact of their CAI on their business. These were, however, not as clearly defined or touched upon by the interviewees as the CAI performance KPIs. The findings at this stage were expressive of a certain difficulty in collecting and analyzing data related to evaluating CAI performance on an enterprise level. The interviews mostly alluded to internal processes being created and developed to improve the collection of data that could be used to measure the impact of the CAI on other customer service channels. The difficulty was expressed regarding both identifying suitable KPIs and making conclusions based on comparisons. It was found that KPIs, where numerical data could easily be collected and analyzed, were favored among the case firms. The KPIs mentioned were: 1) FTE (full-time equivalent) retained, referring to employee worktime saved by the CAI 2) Rate of contacts handled by the CAI (percentage of all contacts handled by CAI) 3) Count of sales leads (how many leads have been collected using the CAI). These are summarized in Table 5. The first-contact resolution was of particular interest to two of the case firms. However, both expressed a difficulty in measuring it as seemingly the only way to collect data on it is through annual qualitative customer surveys. Reduction in CS waiting time was also alluded to as a possible KPI for measuring the CAI's impact. None of the case firms did, however, express having a structured way to determine the CAI's impact on customer waiting time.

It was also found that firms largely base the impact of their CAI on customer satisfaction and value, on assumptions. However, some occasional research into customer experience is conducted through the method of optional questionnaires. This was mentioned by two of the case firms. Questionnaires are largely used to collect data on customers' opinions and habits, but it is hard to specifically single out the CAI's impact in the grand CS scheme. The interviewee from Case B, even expressed that they had given up on looking at end-user feedback ('thumbs-up' and 'thumbs-down') regularly as they had deemed the engagement too low and thus unreliable.

Table 5 KPIs that measure CAI success.

Enterprise level KPIs	**Case A**	**Case B**	**Case C**
First contact resolution	Yes	Yes	N/a
Waiting time	Inconclusive	N/a	Inconclusive
FTE	Inconclusive	Yes	Inconclusive
Rate of contacts handled by the bot (%)	Inconclusive	Yes	N/a
Sales: Leads	N/a	Yes	N/a

2.2.6 Summary of the Client Interview Results

Besides the common objectives, the interviews found that firms consider increasing the efficiency of the customer service function to be one of the main reasons for adopting CAI into their organizations.

In general, the findings from the interviews highlighted that firms' propensity to follow up on their objectives is weak. Pinpointing the CAI's impact is considered hard especially when the effect on other customer service channels is being assessed. The interviews found that firms largely use the number of conversations to measure their CAI's impact.

There were clear differences in evaluation practices and ideology between the 3 different case firms. Case firm B had clear targets set for the rate of contacts they wanted the CAI to handle. This included a clear desire to pinpoint the first contact resolution rate – which they had established as a clear target. Case firm A has an obligation to offer service over the phone and in person, which had a noticeable effect on the prioritization of the CAI channel concerning the other CS channels, also recognized by the interviewees. This dynamic made CAI evaluation crucial as a means to showcase its success. Case A and Case C both emphasized the value of CAI in saving CS agent time. Case B also mentioned using FTE in their evaluation of CAI's impact on other CS channels.

In the effort to support other customer service channels, the CAI's ability to reduce e.g., waiting time on the phone lines was also mentioned in all cases. However, none of the case firms reported that they would be actively looking into the issue making note of the impact the VA has on waiting time. It was thus deemed that how the CAI impacts waiting times in CS is unknown.

2.3 Conclusions from the Literature Review and Client Interviews

This section will summarize the findings of Part 1 of this thesis consisting of the literature review and the multiple case study. The objective of Part 1 was to map CAI objectives and to identify the essential CAI KPIs that contribute to commercial success and performance. The objective was to further define which KPIs are important for Front AI.

The literature review gave a holistic view of the current state of research around CAI evaluation and the approaches most favored in academic literature while the interviews succeeded in collecting a wide range of data on the research topic and establishing a starting point for defining the essential KPIs that Front AI should consider in its development of internal processes around performance reporting and evaluation of CAI.

The findings from the in-depth interviews align with the general findings in the academic literature that the practice of CAI evaluation is still underdeveloped. The case study highlighted that the case firms had clear objectives for their CAI: increasing the efficiency of their customer service function; using CAI as a strategic CS channel (complementary, alongside other CS channels); to increase, promote, and support digital self-service (channels); to diversify the CS offering (24/7 service, text-based communication, CX, and expectations); and to decrease the number of incoming contacts (to all CS channels but especially phone).

The findings allude to a clear task-based approach where whether the chatbot fulfills its intended purpose is the measure of success (Casas et al. 2020, 282-283). However, as the case firms discussed other aspects of success such as user satisfaction and experience, the results indicate that efficiency is not the only determinant of success for the case firms' CAIs. The multi-method approach mentioned in the literature is hence further validated by the collected primary data. Casas et al. (2020) emphasized the popularity of firms measuring the efficiency and user satisfaction of their chatbots which is also evident in the results from the in-depth interviews.

Much like in the literature, the interviews did not yield significant findings on the evaluation of enterprise success or business aspects, except for in Case B where there were measures in place for developing the evaluation of e.g., the rate of contacts they wanted the CAI to handle. However, all case firms mentioned looking at the number of conversations, which according to Peras (2018) Chatbot Evaluation Framework is listed as a metric for measuring business value.

Another finding present in both the primary and secondary data collected was the notion of perceived or assumed performance. De Andrade and Tumelero (2022) interviewed industry professionals on the success and performance of their CAI. The findings indicate that statements were made about the enterprise performance of CAI without real indication of how these gains had been achieved and determined. The same notion was visible in the case interviews with interviewees stating that their CAI performed well on user satisfaction without really having the measures in place to evaluate it. This finding is a further testament to the lack of practice-based knowledge in CAI evaluation.

What is evident in both the literature and practice is the lack of distinction between metrics that contribute to the evaluation of enterprise success or "higher level goals" such as generating business value, and metrics that focus on measuring the CAI's performance on a more technical level. Both literature and practice have heavily focused on looking at the more minute details of CAI performance such as the prediction, the number of times the answer is correct, and whether the response is appropriate. While some of these metrics might give an indication of the CAI's overall success – such as number or conversations – it should be emphasized that they only yield results of enterprise success if put into context or compared. In this example, the rate of contacts the CAI handles is more descriptive of the value the CAI produces as it is compared to other customer service channels. The objectives mentioned by the case firms all emphasize CAI's role in customer service. The aim is to use CAI to diversify, promote, improve, and increase the efficiency of other customer service channels. This being the case, firms should focus on enterprise-level KPIs that measure the success of these targets i.e., not only the CAI's performance independently but also in relation to the existing customer service functions and channels.

While putting CAI performance in context is necessary it is still a wildly underdeveloped area in both research and practice. The first step, however, is to understand the objective of the chatbot and to determine which metrics should be used to assess its success. As Front AI's customers tend to deploy chatbots with the objectives of improving the efficiency of their customer service function – as highlighted in the findings – the KPIs most relevant to Front AI are the ones that aim to measure the performance of these objectives.

Peras' (2018) Chatbot evaluation framework (CEF) gives a comprehensive view of KPIs. Peras (2018) emphasizes the adoption of any or all perspectives depending on the CAI's objective. To pinpoint the most suitable KPIs and metrics for Front AI, Peras' (2018) CEF should be modified according to the objectives mentioned by Front AI's clients in the interviews, highlighting the important perspectives.

The perspectives and categories in Peras' (2018) Chatbot evaluation framework, most suited for Front AI to look at in evaluating CAI are: User experience perspective (Usability, Performance, Satisfaction), Information retrevial perspective (Accuracy, Efficiency), Technology perspective (Humanity), Business perspective (Business value). These are summarized in Figure 1.

Figure 1 Perspectives and categories for CAI evaluation at Front AI based on Peras' (2018) Chatbot evaluation framework (CEF). (Illustration: Hedvall 2023)

The selected perspectives assure a comprehensive evaluation of the type of CAI commonly deployed by Front AI's clients. To measure these aspects of CAI the thesis proposes the following KPIs, and metrics, found in Table 6, based on the findings from the literature review and client interviews. The categorization of metrics takes inspiration from Peras' (2018) CEF table and is supplemented with other KPIs mentioned in the literature and client interviews.

Table 6 KPIs and metrics for Front AI for evaluating CAI success and performance.

User experience perspective	Usability, Performance, Satisfaction	• Count and rate of successfully solved conversations (in-scope resolution rate) (Peras 2018; De Andrade & Tumelero 2022) • Duration of conversation (Pera 2018) • Turn-level user satisfaction ('thumbs-up' and 'thumbs-down') (Interviews) • Resoluteness (Rate of contacts handled by the CAI) (De Andrade & Tumelero 2022)
Information retrieval perspective	Accuracy, Efficiency	• Number of intents (Interviews) • Most popular intents (Interviews) • Unknown message count (Interviews; Peras 2018) • Average count of turns per task (Lewandowski et al. 2023; Peras 2018) • Human-handover rate (Lewandowski et al. 2023)
Technology perspective	Humanity	• Rate of fallback responses or errors (Peras 2018) • AI model success score (Casas et al. 2020)
Business perspective	Business value	• Conversation count (Interviews; Peras 2018) • Duration of conversation (Pera 2018) • Engagement (average time spent on the website when interacting with CAI) (Jadeja & Varia 2017) • Resolution rate (overall rate of successful conversation) (De Andrade & Tumelero 2022) • Adoption rate (De Andrade & Tumelero 2022) • Automation rate (De Andrade & Tumelero 2022) • Lead generation (Interviews) • Average waiting time (Interviews) • Volume Curve Flattening (Interviews, De Andrade & Tumelero, 2022) • Availability (Interviews) • FTE (Interviews; Osei-Mensah et al. 2023; De Andrade & Tumelero 2022) • Estimated latency time (De Andrade & Tumelero 2022) • First contact resolution (Interviews) • Number of repeated contacts (Peras 2018)

The assessment and inclusion of KPIs and metrics for each corresponding perspective in the recommended standard is subjective and limited to the findings collected for this thesis and the author's understanding of the topic area. The recommended assortment should also be complemented with other KPIs that Front AI may find of significance. While this collection presents a tailored CAI evaluation standard for Font AI, the approach and KPIs should be adjusted based on the objective of the specific CAI being evaluated –which might vary significantly among Front AI clients. However, this standard can be used as the basis for understanding the basic KPIs that are significant for Front AI and consequently be used for creating more comprehensive versions of such a framework.

The framework purposefully outlines a wider evaluation of User satisfaction such as Chatbot interface and personality evaluation as this is not emphasized in the client interviews to the extent that it would be considered significant. However, Case B displayed a strong attachment internally to their CAI which had been equipped with a fully formed persona. This reportedly increased the use of the CAI, also internally increasing awareness of the opportunities and strengthening the chatbot's position within the organization. The literature further emphasizes the importance of user satisfaction especially in commercial CAIs highlighted by Lewandowski et al.'s (2023) study, which defines the success of a commercial CAI to be solely determined by user satisfaction. While a comprehensive evaluation of aspects related to user experience is warranted, it is not in the scope of the regular maintenance that Front AI engages in. As the aim of this thesis is to contribute to the development of Front AI's internal processes around performance monitoring of certain aspects of their client's CAI chatbots with a focus on evaluating CAI impact on an enterprise level, the specific evaluation of user experience is deemed to be out of scope by the commissioner.

However, the results from the theoretical part do highlight other findings on CAI evaluation worth considering in contexts outside of the scope of this thesis, such as the emphasis on chatbot privacy and security concerns that affect user experience. These should be considered in the long-term evaluation and development of chatbots. For conducting regular monitoring and maintenance these aspects are less emphasized. Other aspects – mentioned in both literature and the interviews – that can influence enterprise success can be for example prediction which refers to how data collected from CAI can help the organization predict customers' intentions and thus help to improve the user experience and progressively the customer service (De Andrade & Tumelero 2022). The interviewees also highlighted the effects that easy adoption and ease of use of the CAI platform can have on the success of the initiative. While these aspects provide clear benefits for firms, they are often overlooked due to them being harder to measure and pinpoint. These "soft benefits" such as prediction of user intent, ease-of-use of the solution, and proactiveness were mentioned as something highly valuable by the interviewees, as they can help firms react quickly to changing service landscapes. This in turn can have a positive effect on multiple aspects of CAI enterprise success e.g., engagement, estimated latency, and waiting time which in turn creates business value.

3. PART 2: DEVELOPING A CAI EVALUATION PROTOTYPE FOR FRONT AI

In the previous chapter, the thesis presented a collection of KPIs for CAI evaluation tailored to Front AI's need to establish an understanding of KPIs that can be used in evaluating CAI success and performance on an enterprise level. The suggested collection comprises CAI KPIs and objectives identified in academic literature and interviews with Front AI clients.

It is however evident that the KPIs and objectives themselves – although now defined – do not add direct value to Front AI. To fulfill the purpose and achieve the aim of this thesis, the information on CAI success and evaluation needs to be packaged in a way that will enhance knowledge sharing and the cultivation of the necessary skills for CAI evaluation within Front AI. As the findings should support the development of internal processes around CAI evaluation and performance reporting, they should be communicated in a way that will support their adoption and contribute to knowledge acquisition and learning around the topic. Further elaboration is needed so that notions of CAI success, jectives, and KPIs can be adopted and utilized within Front AI, aligning activities efficiently and enabling data-driven decision-making.

The objective of this part of the thesis is to develop a prototype – using a design thinking approach and service design tools – through which the collected KPIs and objectives can be communicated to the users at Front AI. The aim is to satisfy the second objective of the thesis, using service design methods and a design thinking approach to create a user-centric solution. This part of the thesis thus focuses on the case at hand and the design process around the development of the prototype, taking into account the current context in which the newly discovered information will be adopted.

The design thinking approach is a methodology used to solve complex problems (Dam 2022). It is usually described as a non-linear, iterative process with multiple phases ranging from 4-7. A famous design thinking model developed by the Hasso Plattner Institute of Design at Stanford (d.school) presented by Dam (2022) is the five-stage design thinking process (Figure 2) with the following stages: 1) empathize, 2) define, 3) ideate, 4) prototype, 5) test.

This design thinking model (Dam 2022) was used to guide the activities in the development process of the prototype. Some of the stages of the design process were conducted concurrently with some earlier stages occurring repeatedly as more information was uncovered. The following sections will present the design process starting with the empathize phase, introducing the investigation into the end users of the final prototype. After that the process of defining the design problem will be introduced after which the thesis will continue with presenting the ideation phase, the prototype phase, and finally the testing phase.

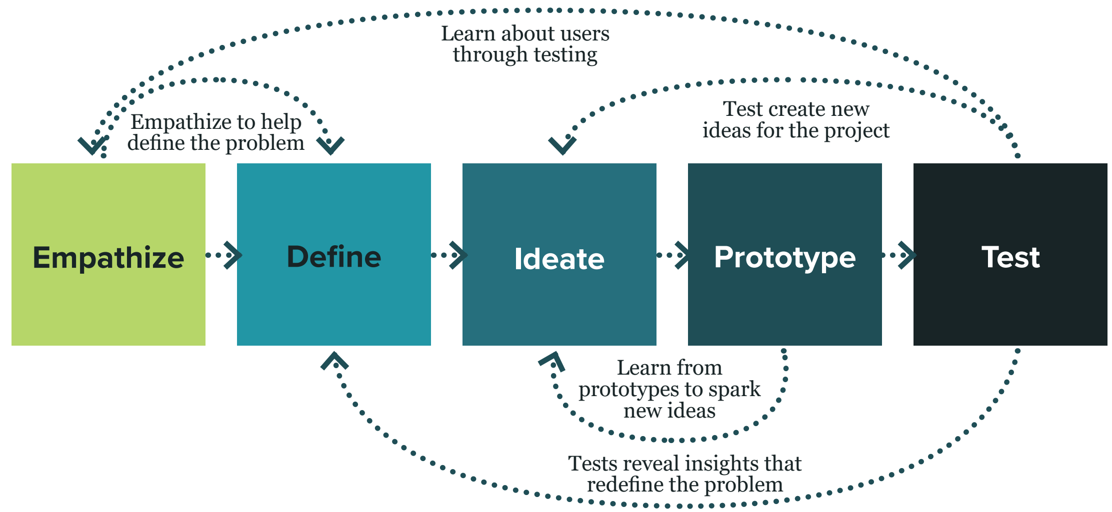

Figure 2 The design thinking process developed by Hasso Plattner Institute of Design at Stanford (d.school). (Interaction Design Foundation 2022, edited by Anna Hedvall)

3.1 Case Background

Front AI is in the process of developing its business and knowledge management processes with an emphasis on efforts focused on the development of the CAI evaluation and performance reporting practices. Reporting on CAI performance is a part of the services offered to clients and is conducted every month by AI Trainers. The reporting involves the monitoring and evaluation of specific aspects of deployed CAI implementations.

Front AI has in their previous development of new business and knowledge management processes around reporting focused on creating tools that make the process more efficient and streamlined. While the processes and tools have improved, the need for support in performing reporting still exists at Front AI. As new information on CAI success and evaluation gets introduced, demand for learning and support increases. Front AI's main challenge has thus centered around how to convey information that AI Trainers need to do their job successfully, providing and facilitating sufficient learning opportunities.

3.2 Empathize: User Research

Understanding the end users' true needs and wants is a critical element of design thinking. The final ideas and solutions should be grounded in a deep empathetic understanding of the target user. (SocialUP 2023b.)

This section will present the user research for the productive part of the thesis with Front AI as the case participant. The section will describe the process of gaining user insight through data collection and analysis using service design methods. The process began with collecting user data through contextual interviews. The information was then analyzed using service design methods such as affinity mapping. Based on the gathered data, a persona was created, with the help of an empathy map canvas, to illustrate the findings of the user research and ensure appropriate focus throughout the entire design process.

3.2.1 Understanding the End User: Contextual Interviews

Contextual interviews are a qualitative research method used to understand the target user better. Contextual interviews provide an understanding of the needs, emotions, and expectations of the user. Furthermore, the method provides information about the environment the user acts in, which is useful for creating personas. Contextual interviews allow the user to demonstrate actions or thoughts in detail and in context, which is useful for understanding particular experiences. (Stickdorn et al. 2018, 20.)

Contextual interviews can be conducted rather openly, following a leading research question, or in a semi-structured way (Stickdorn et al. 2018, 20). The interviewer should attempt to ask the interviewees about specific experiences and for them to demonstrate details of these concrete experiences, as it is often easier to articulate complex experiences by referring to concrete examples (Stickdorn et al. 2018, 20). The interviewee can use props or artifacts to articulate problems and needs, to help provide an understanding of the situational context (Stickdorn et al. 2018, 21). Contextual interviews also allow the interviewer the opportunity to observe the environment the user acts in while also making the interview more engaging and open as the environment is familiar to the interviewee (Stickdorn et al. 2018, 21). As interviewees tend to remember more specific details when they act in an environment and context familiar to them, the interviewer also gains a more holistic understanding of the user (Stickdorn et al. 2018, 21).

The purpose of the user research in the Front AI case is to understand the perceived issues and needs of the Front AI employee. It is essential for the design process and development of the final solution, that it reflects and aligns with the end user's needs. The second objective of the thesis being that of the development of a prototype to support internal reporting processes that will facilitate the understanding of CAI success and KPIs, prerequisites an understanding of the current situation and issues that the Front AI employees face. The thesis will focus on a particular Front AI employee group – the AI Trainers – as they are the main beneficiaries and responsible for conducting reporting and CAI evaluation. The AI Trainers play a key role in Front AI's operations and as the responsibilities of an AI Trainer require proficiency in all areas of CAI, they are the primary target users for the prototype solution.

The plan was to map the AI Trainers' knowledge and understanding of concepts related to CAI success and KPIs using an existing tool. The final sample, consisting of 7 interviewees, was requested to fill in the information regarding a certain client case and define its objectives, goals, and KPIs in advance of the interviews. The results were then assessed in a contextual interview where the participants were allowed to reflect on their process, ask questions, and brainstorm out loud to explain their reasoning. As the interviews used real client cases as examples, they are confidential and will thus only be discussed on the detail of what data about the target user's approach, issues, feelings, needs, and wants was collected.

The contextual interviews were conducted over virtual meetings where the interviewees had the task of presenting the information they had gathered in the tool regarding the client's case. The interviews took a semi-structured approach with some interview guidelines established beforehand to ensure proper focus. The research question What is the current understanding of CAI's success, objectives, and KPIs at Front AI? was used as a leading question as recommended by Stickdorn et al. (2018, 21). The interviews were conducted with the help of a Front AI stakeholder in charge of the tool. This allowed for the thought process and dialog between the interviewee and the stakeholder to be observed by the interviewer. This gave a holistic view and allowed the interviewee to freely ask questions and express their experiences with the help of the tool, according to the guidelines of a contextual interview (Stickdorn et al. 2018, 21).

Data from the interviews was collected in the form of interview notes due to the confidential nature of the client cases being discussed. The contextual interviews yielded findings on the general knowledge and understanding that the target user has about CAI success, objectives, and KPIs. These will be explored in the following sections.

3.2.2 Analysis of the User Research: Affinity Map

To analyze the data collected during the contextual interviews theming and an affinity map was used.

The interview data in the interview notes were analyzed and themed. Theming is a method used to classify and distill research data. The method is used for finding commonalities, deviations, and connections in the data. It is a form of data processing that helps with the defining and scoping of the design problem, creating a good prerequisite for the ideation of a solution. (Kurronen et al. 2015, 22.)

The theming was conducted using an affinity map (SocialUP 2023a), affinity diagram (Elmansy 2023), or clustering map (SDT 2023a) which is a visual display of the theming process. The method helps in building connections between data fragments and to build an understanding of the relationship between e.g., separate interviews (Elmansy 2023). The tool aims to organize data and findings into groups in an interrelationship diagram (Elmansy 2023). Affinity diagrams are especially useful in displaying research outcomes in a structured way (SDT 2023a).

The interview outcomes were analyzed using templates and features in the Miro -software program as opposed to the standard way of using paper or sticky notes on a whiteboard. As Miro carries similar features – only digitally – it was deemed appropriate. Furthermore, as the interview notes were digitally collected, direct quotes from the interviews could easily be compiled, analyzed, and clustered using Miro. The software also provides the opportunity to make reassessments of the clusters and groupings, with ease, which makes the analysis of the findings convenient.

The sorting of the contextual interview data began with compiling all interview responses and notes into tables using a pre-existing template: User interviews to research insights. Each separate interviewee's responses were compiled into its own table. After the responses had been documented – each response on its own virtual sticky note – they were ready to be grouped. The grouping using the affinity map approach began by sorting each sticky note based on whether the response displayed or indicated a certain behavior or feeling. As the purpose was to build a holistic understanding of the user, understanding what motivates them or what is challenging to them, was of interest. After each sticky note had been sorted as either something the end user does or thinks, the sorted sticky notes were assessed for similarities. The responses from each of the 7 interviews were compiled and sorted into sub-groups, producing findings on the connections in the data. The outcome of the affinity mapping produced 4 themes explaining behavior and 6 themes explaining thoughts and feelings about CAI evaluation. These were then distilled into 10 concluding remarks to communicate the findings from the contextual interviews. Each remark had quotes from the raw interview data to support the argument. The concluding remarks from the contextual interviews were then ready to be used in the process of creating a user persona.

3.2.3 Presenting the Findings from the User Research: User Persona and Empathy Map Canvas

User personas are a tool frequently used in design thinking and in service design to understand the end user's needs and wants. It is a tool for visualizing, synthesizing, and analyzing data (Stickdorn et al. 2018, 51). User personas determine the target or end user of the outcome (Kurronen et al. 2015, 31) helping convey what success for them will look like (SocialUP 2023b). A User persona is a fictitious and archetypal character that represents a group of end users and should be based on data from real users, to avoid stereotypes and biases (Miaskiewicz & Kozar 2011).

User personas can be used as a tool for unpacking user research so that observations from e.g., interviews are added and visualized in a user persona -template (Kurronen et al. 2015). The objective is to encapsulate characteristics of the end user that illustrate their needs, wants, behavior, and real-world perceptions of a certain issue (Stickdorn & Schneider 2012). These characteristics should be derived from research data and information that is not indicative of the user's behavior as information related to demographic and socio-economic aspects can be misleading resulting in a user persona built on assumptions and stereotypes (Stickdorn et al. 2018, 51).

The goal of the empathy map is to build empathy with a specific user. Empathizing with the user is important when developing a solution that users interact with as it helps the team understand the user on a deeper level, becoming aware of their real needs. Empathy maps are usually used to supplement the user persona with characteristics beyond goals, skills, and interests. Empathy maps can therefore be used to describe personas. (Ferreira et al. 2015.)

The empathy map method has been developed over time with the first and perhaps most standard versions consisting of four different areas: hear, think & feel, see, say & do. The canvas has since been complemented with additional areas such as pain and gain. (Ferreira et al. 2015.) There have been multiple versions of the empathy map over the years and recently a broader take on the empathy map has surfaced with the introduction of the Empathy Map Canvas by Gray (2017).

While any empathy map can be used to improve on the findings from user research –deepening our understanding of our target user defined in the persona – Ferreira et al. (2015) describe in their paper how the process of creating a persona can also be made easier by using an empathy map.

The process of displaying the findings from the analysis of the user research for this thesis started with the use of the Empathy Map Canvas. Much alike to the findings of Ferreira et al. (2015), the process of creating and describing the persona felt easier when approaching it through the Empathy Map Canvas. The filled in Empathy Map Canvas is displayed in Figure 3.

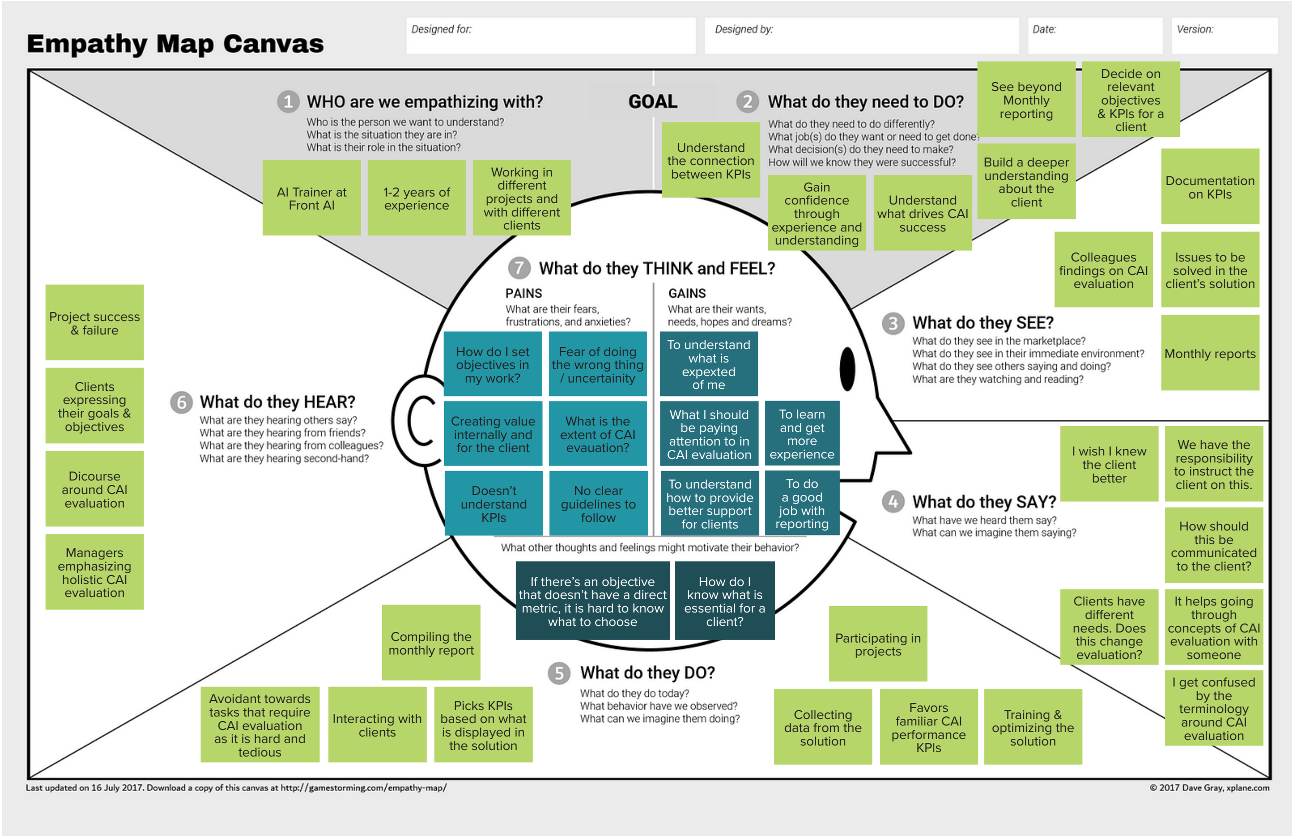

Figure 3 The Empathy Map Canvas by Gray (2017) filled in with the findings from the contextual interviews. (Gray 2017, modified by Hedvall)

The process of creating the persona started with a review of the concluding remarks from the affinity map which were then modified and fitted into their respective attribute on the Empathy Map Canvas with conclusions pertaining to e.g., actions, modified into statements under the "what do they do?" section of the persona, etc. As the findings from the interviews were focused on behaviors or feelings it was easier to use the Empathy Map Canvas to categorize them. The empathy map canvas was then used in the creation of the persona, with the findings being distilled into the behaviors and characteristics of the fictional user persona. The final product of the Persona is pictured in Figure 4.

A conscious choice was made to have the persona focus only on the specific context of CAI evaluation, the goals and pain points focusing on what the user feels and thinks when they are required to partake in CAI evaluation. In other words, it was important to present the persona from the angle of what habits they have, what they find challenging, and what their needs are when it comes to CAI evaluation. For this outcome, it was deemed best to use both the empathy map canvas and persona in the design process. This in turn ensured that during the development of the prototype optimal user focus persisted.

In the process of creating the persona and filling in the empathy map canvas, bias was limited by focusing on insights from the user research data and not speculation. Data that did not cause a certain behavior was not included to avoid making assumptions based on stereotypes.

Jessie

About Jessie

Jessie is an AI Trainer with 1 year of experience working at Front AI. Jessie is a good problem solver that has learned to identifying model strengths and weaknesses and is known for being able to break down complex concepts into understandable lessons for Front AI clients. Jessie finds their work to be rewarding but sometimes hard as it can lack structure and clear expectations.

Demographics

AI Trainer
Works remotely
Works digitally

Behaviors & habits

- Interacts with clients in meetings and in chats
- Works together with colleagues on client projects
- Puts off CAI evaluation as it is tedious, repetitive and complicated
- Turns to exsisting information and documentation to find answers
- Analyzes model statistics and compiles maintenance reports
- Trains and optimizes AI models
- Turns to colleagues for help and inspiration around CAI evaluation
- Favors familiar CAI objectives and KPIs

Needs & Goals

- Understand what drives CAI success
- Contribute to making the process of monthly reporting better
- Gain confidence through experience and learning
- To understand what is expected ofz them
- To learn and get more experience
- Decide on the relevant objectives and related KPIs for a specific client case
- To do a good job with the reporting
- To understand how to provide better support for clients
- Empathize with the client (understand their objectives and seek solutions to measure them)

Pain points & frustrations

- Not knowing the client and th
- Uncertain of what they should focus on in their reporting. How far should I go? What is expected of me?
- Fear of doing the wrong thing or making the wrong decision
- No firm grasp of the KPIs or the extent of CAI evaluation.
- Missing clear guidelines or structure in setting objectives for CAI success

Attribute scales

Work experience
1 2 3 4 5

Motivation
1 2 3 4 5

Technical familiarity
1 2 3 4 5

Decisiveness
1 2 3 4 5

Early adopter
1 2 3 4 5

Figure 4 The Persona: Jessie. (Illustration: Hedvall 2023)

3.2.4 Summary of the User Research in the Case Front AI

In this chapter, the thesis explored the current situation and understanding at Front AI around CAI evaluation using a design thinking approach. A target user was selected, and contextual interviews were conducted for the purpose of creating an understanding of the end user. To ensure a user-centric ideation and design process, the thesis set out to empathize with the end users. Service design tools such as the Contextual Interview, Affinity map, Empathy Map Canvas, and Persona were used to collect, analyze, and highlight the results and findings from the user research in the case of Front AI.

3.3 Case Front AI: Defining the Design Problem

Defining which user problem to solve is an essential stage in the design thinking and service design process. This stage usually comes after an understanding of the end user has been established. The define stage is for synthesizing the information that we have gathered and analyzed during the user research phase. Synthesizing involves the process of organizing, interpreting, and making sense of the data that has been gathered to create a problem statement. (Dam & Siang 2019.) The purpose of this is to get a clear idea of what problem we are trying to solve for our end user (Stevens 2019).

Thoroughly developing key insights and defining the design problem will help create a clear objective for the development of the solution (Stickdorn et al. 2018, 60). A clearly stated objective will also ensure appropriate focus and that the solution that is being developed is aimed at solving the problem in question (Stevens 2019). Key insights and problem statements can be used later to evaluate ideas, concepts, and prototypes (Stickdorn et al. 2018, 60).

3.3.1 Developing Key Insights from the User Research and Defining the Design Problem

The development of key insights started by inspecting the findings from the user persona and Empathy Map Canvas. The process began by identifying patterns in the user data displayed on both canvases. Thereafter a template introduced by Stickdorn et al. (2018, 60) was used to define the key insights. As the persona had been structured around the behaviors, habits, needs, goals, pain points, and frustrations, it was easy to identify the user's perceived problems. After the key insights had been comprised the design problem was defined.

The identified key insights:

- Jessie analyzes model statistics and compiles maintenance reports **because** they want to do a good job reporting on CAI performance **but** Jessie is uncertain of what they should include in the report and what KPIs they should use.
- Jessie trains and optimizes client AI models **because** they want to add value to clients and provide a good service for them **but** still feel that they don't know the client and their objectives, needs, and wants well enough.
- Jessie works together with colleagues on client projects **because** they will feel more confident when they have gathered experience and knowledge **but** still fear failure, doing the wrong thing, or making decisions independently.
- Jessie interacts with clients in meetings **because** they seek to understand the client's objectives **but** Jessie lacks a firm grasp of how to measure these, the KPIs to use, and what constitutes CAI success.
- Jessie avoids exploring unfamiliar CAI metrics **because** they are uncertain of their meaning and significance and whether it is something Jessie should be focusing on **but** Jessie feels there are no clear guidelines for what is expected of them as an AI Trainer.
- Jessie turns to existing information to find answers **because** Jessie wants to determine what the relevant objectives and KPIs are for a specific client **but** there is no sufficient documentation on CAI success and setting objectives.

The key insights highlighted a common issue among the AI Trainers at Front AI, the general finding indicating that AI Trainers struggle with CAI evaluation due to lack of experience, guidance, and set expectations. AI Trainers who were frequently conducting reporting, CAI evaluation, and performance monitoring felt as if they did not have the necessary skills or information to do what was expected of them. They also expressed how determining objectives and KPIs for clients is hard, as understanding CAI success on a holistic level is less emphasized in the work of AI Trainers. The following design problem statement was thus identified:

The AI Trainers at Front AI need a way to feel confident and supported in the process of defining KPIs and setting measurable objectives for client CAI, as they regularly conduct evaluations and performance reporting that make them feel uncertain.

In addition to the defined design problem, design drivers were established to aid in the development of the prototype and to guide the design process.

3.3.2 Design Drivers

Design drivers are pieces of information identified during the user research, that guide the ideation and design process. Well-defined design drivers help with developing clear and user-centric concepts. By using design drivers designers can ensure that the end user's needs and desires are front and center in the design process. (Lahden ammattikorkeakoulu & ProAgria Etelä-Suomi ry, 2018.)

Design drivers help capture what a well-executed concept might feel like while also providing a point of reference that helps us judge an evolving design's viability. The use of design drivers felt like the right tool to enrich the clearly defined design problem as the method helps with setting the aspirational goals and requirements for the concept. It also provided the design process with tangible aims with which success could easily be assessed. (Driver 2018.)

- Informative but efficient
- Familiar and predictable
- Adjustable and modifiable

The final concept should be informative, in the sense that it should provide the end user with the necessary information to define CAI objectives and KPIs. The concept should give out sufficient information to help the end user make an educated decision. While conveying information should be the main feature of the concept, it should not be overflowing with information, as this might lead the end user to feel overwhelmed as they are faced with information overload. The concept should thus present information efficiently and leave out any unnecessary information. Only the most central pieces of information should be included in the concept.

The concept should also feel familiar to the end user. The point of the concept should not be to learn a new way of doing something but to learn something new, thus it should follow a familiar logic. For something to be familiar to the AI Trainer, it should be presented, documented, and communicated using the same methods as used previously at Front AI. The concept should prioritize methods and tools that are familiar to the end user as this will also create predictability – the notion of knowing what to expect. The concept should avoid putting the end user in a position where they need to deduce or experiment with the tool to get a result.

Finally, the concept should be easily adjustable and modifiable. While the concept is developed for the AI Trainers at Front AI, it should still allow for flexibility. The concept should allow for further expansion and iteration based on Front AI's and the AI Trainer's changing needs and aims. It should be easy to add or change the content of the concept.

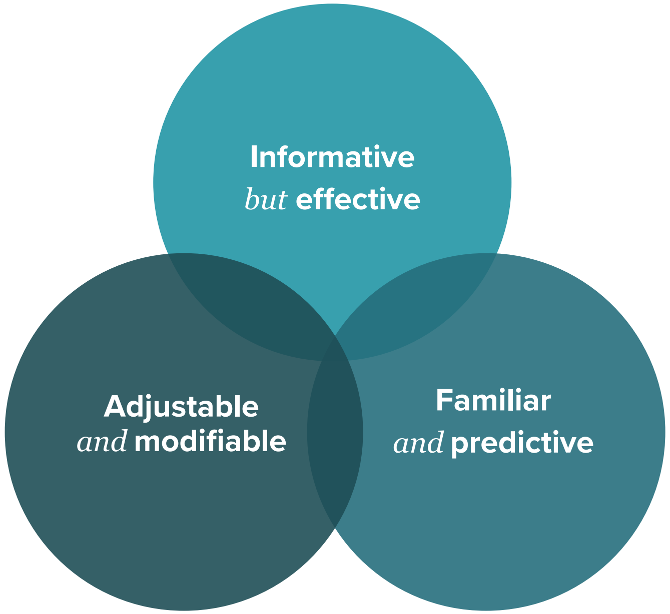

Figure 5 The design drivers for the prototype. (Illustration: Hedvall 2023)

3.4 Ideation

After the design problem had been defined and design drivers established the process moved onto the ideation phase. The ideation phase comprises the generation of ideas and solutions to the design problem. The ideation phase allows for looking at the problem from different perspectives and creating innovative solutions using different techniques and service design tools. (Dam 2022.)

The ideation began with a review of the design problem statement and the identified objectives and KPIs presented in Part 1 of the thesis. Initial ideas were quickly prototyped and brainstormed but they were deemed to be too biased and reliant on preconceived notions of what the solution should be. To counter this, another approach was deemed to be needed to get rid of bias, to stimulate free thinking, and to ensure the solution was rooted in the defined design problem.

3.4.1 Customer Journey Map and Mindmap

The ideation was conducted in Miro as data from the user research had already been gathered and showcased there, which allowed for easy referencing to the user persona and empathy map canvas. At this point in the ideation, an additional service design tool and canvas, the Customer Journey Map, was deployed to help expand the problem space. A Customer Journey Map (CJM) was created for the Persona Jessie. The map was created to illustrate the process and scenario of selecting CAI objectives and KPIs using the existing tool created for this purpose and also used in the contextual interviews.

Customer Journey Maps are synthetic representations describing step-by-step how a user acts with a service, organization, or product (SDT 2023b). CJMs describe, from the perspective of users, what happens at each stage of the interaction, displaying touchpoints, relationships, obstacles, and barriers as well as which positive or negative emotions the user experiences during the interaction (SDT 2023b). The CJM is used for example in instances where there is a need to see how user experiences meet user expectations and when there is a need to improve ideas and designs (Interaction Design Foundation 2019).

Once, the CJM had been created, to represent the persona's journey through the process of selecting CAI objectives and KPIs, the filled canvas was analyzed. The analysis focused on identifying the crucial touchpoints and steps in the process that caused the most negative experience for the user. Two particular pain points were identified, with one being the filling of the tool and the other one being the looking for additional information to aid in the filling of the tool. These were the steps in the process that caused the worst experiences for the user. These were also concrete steps that could and should be improved on, directly related to the core issue of the end user defined in the design problem statement.

Using the two pain points as a reference, the generation of ideas resumed. The CJM allowed for a more focused approach to the ideation, and the generation of concrete ideas to the different jobs-to-be-done, touchpoints, pains, and gains in the CJM, using a "How might we...?" approach (Stickdorn et al. 2018, 83). Sticky notes with perceptions and ideas were attached to the CJM to highlight what could make e.g., a certain touchpoint better. A snapshot from this process is dispayed in Figure 6.

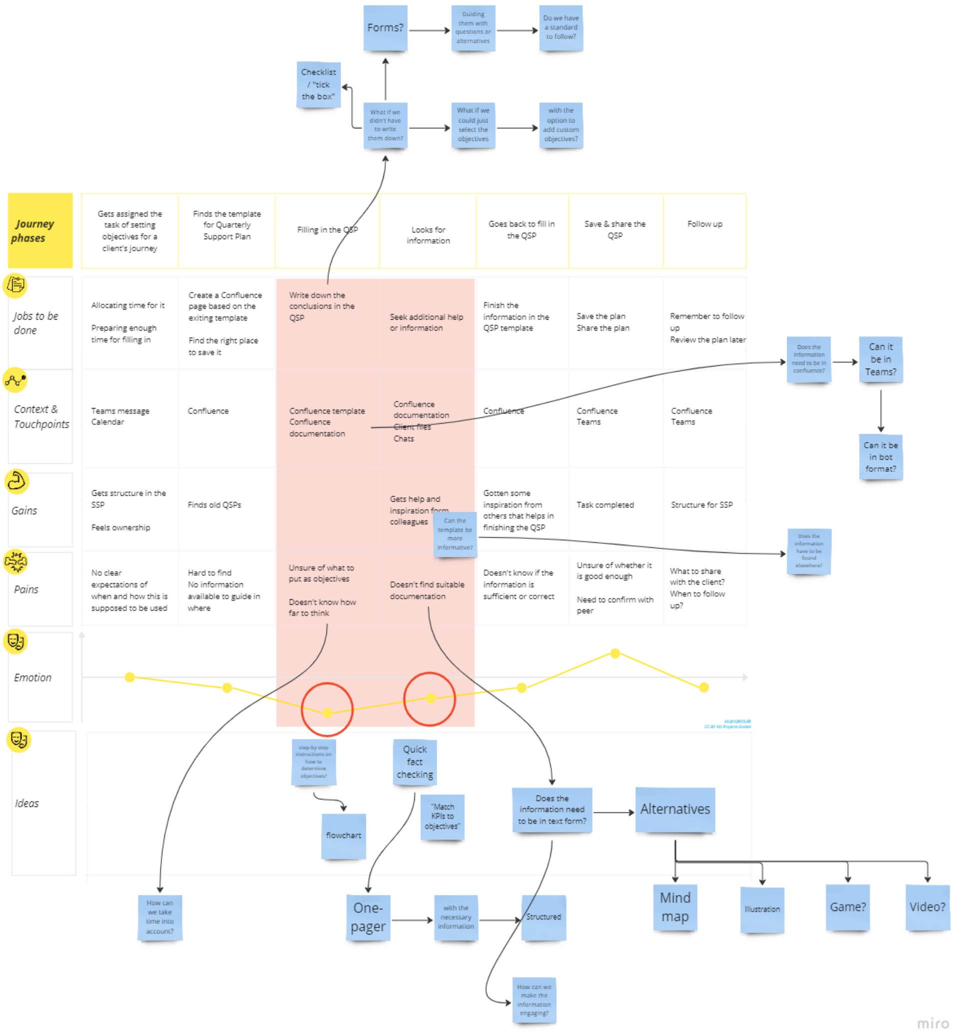

Figure 6 A snapshot of The Customer Journey Map with ideas and pain points highlighted in red. (Illustration: Hedvall 2023)

The ideas, in sticky note format, were then compiled into a mindmap. Mindmaps are a tool that helps structure thoughts around specific topics (STD, 2023). They are also good for fostering unconstrained thinking and uncovering connections between ideas (SDT, 2023c). The mindmap allowed for the expansion and laddering of ideas in addition to highlighting similarities, relationships, and connections between the ideas. This approach allowed for even more blue ocean ideas to be considered and expanded on.

After all ideas had been exhausted using the mindmap they were assessed based on how well they reflected the design problem. The design drivers were also used to categorize and rate the ideas' viability and suitability to solve the problem. Figure 7 displays the mindmap. At this stage in the process, the most attractive ideas were selected for the prototyping phase.

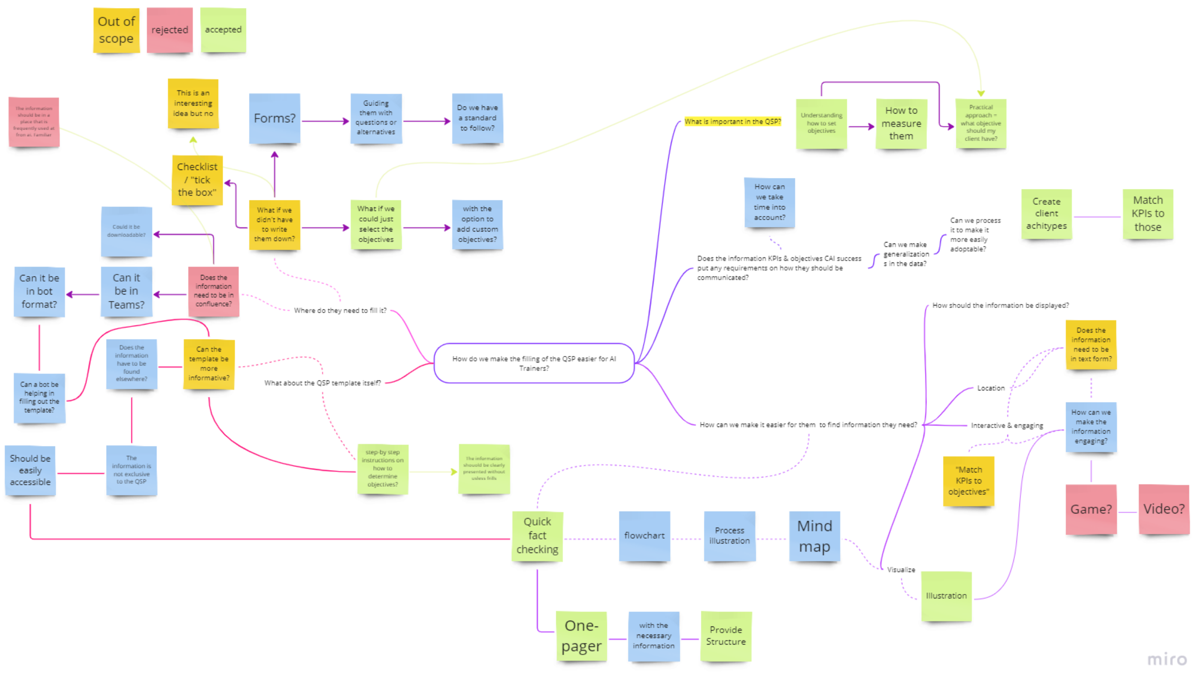

Figure 7 The mindmap with the ideas marked as either approved, rejected, out-of-scope or other. (Illustration: Hedvall 2023)

3.5 Prototyping and Testing

Prototyping is the process of producing several scaled-down versions of the solution to investigate the ideas generated in the ideation phase of the design process. The prototyping phase is an experimental phase that aims to identify the best solutions for the design problem. The prototypes can be tested on, for example, a focus group consisting of end users and accepted, improved, or rejected based on what the end users think of them and how they find the experience of the interaction. The goal of the prototype phase is to get a better idea of the solution's limitations and what the users feel when they interact with it. (Dam 2022.)

The ideation phase highlighted the need for information on CAI objectives and KPIs as, in the process of deciding on objectives and KPIs and looking for information regarding these, the lack of information was a clear pain point for the end user. The approved ideas that would be made into prototypes, were thus centered around how to best convey this information. While the information on how to set CAI objectives and KPIs is not exclusive to just the tool, it was clear that the stage of filling in the tool was perceived as the most difficult one for the end user. So instead of focusing on making a general guide on CAI objectives and KPIs, it made sense to create a guide for this specific task that the end user needs to perform. This thinking was further supported by the design problem statement that emphasized the AI Trainer's need to feel "confident and supported" in the process of choosing CAI objectives and KPIs. Thus, a tool for helping pick objectives would give more value than a general information sheet.

The ideation phase also included brainstorming around the different ways in which information can be conveyed, with everything from illustrations, audiovisual material, and games being explored. However, the design drivers helped with scoping the best options in this scenario, emphasizing the need for the solution to be familiar and predictable to the end user, marking off the more extreme and engaging options. The ideas of the different ways to display information were also assessed based on attributes such as understandability, forgiveness, and affordance to ensure that the end user would not have to learn how to use the solution but instead use the solution to learn how to set CAI objectives and KPIs. The need for the solution to be easily modifiable and adjustable was also considered in the ideation of the solution.

A couple of different ways to convey information were explored through quick prototyping in Miro. Some light benchmarking around information design was done to get a sense of what the prototype could look like. A quick prototype based on a table format was developed and selected for testing on the end users.

3.5.1 Testing the Quick Prototype

The testing phase is the final phase of the design thinking process. However, as the design thinking process is iterative, the results of the testing are often used to discover problems or redefine the prototype leading the designer to return to previous phases in the design thinking process. Testing prototypes is essential to understanding how the end users think, behave, and feel about the solution that has been created. Testing will allow for refining the solution and exploring alternatives, based on the end user's experience. (Dam 2022)

The testing of the quick prototype was conducted in person with two AI Trainers from Front AI. The testing took one hour, and the participants were able to view, interact, and discuss the quick prototype. The testers were randomly selected from the pool of available AI Trainers and asked to explore the prototype displayed in Miro. The AI Trainers gave feedback on the look and feel of the prototype, what worked and what did not work, and how the prototype helped them in the task of defining CAI objectives and KPIs. Snapshots from the testing session are displayed in Figure 8 and Figure 9. The testing was conducted in an unstructured way and notes on areas of improvement were added in real time in Miro.

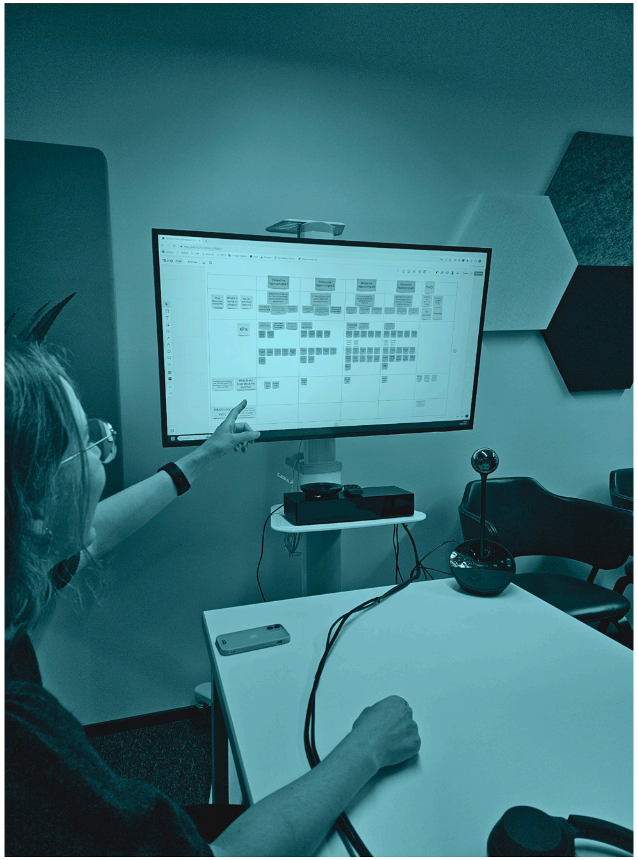

Figure 8 AI Trainer exploring the quick prototype in Miro. (Image: Hedvall 2023)

Figure 9 The AI Trainers are giving feedback on the look and feel of the prototype. (Image: Hedvall 2023)

3.5.2 Ideation and Prototyping Based on the Guick Prototype Test Results

The testing yielded valuable findings on what the AI Trainers thought about the prototype, highlighting the importance of clear call-to-actions, structure, visual clues, and wording. The general findings from the testing were that the AI Trainers found the prototype to help them think on the topic of CAI success in a way they had not thought of before and that the prototype helped them in the task of choosing objectives and KPIs for CAI, despite still feeling the need to get a second opinion.

After the end user test, the results of the testing were compiled and a new round of ideation and "How might we...?" was conducted to generate ideas based on the newly discovered issues. Prompts were derived to guide the adjustment and modification of the prototype. At this stage, other alternative approaches were explored and prototyped regarding the information design but rejected as non-viable options due to being inferior in terms of clarity and understandability compared to the quick prototype.

The final round of prototyping based on the test results, yielded a prototype solution that filled the requirements and purpose of the design brief. Some early iterations of the final prototype are pictured in Figure 10. The final prototype was then taken for another round of testing, described in the next section.

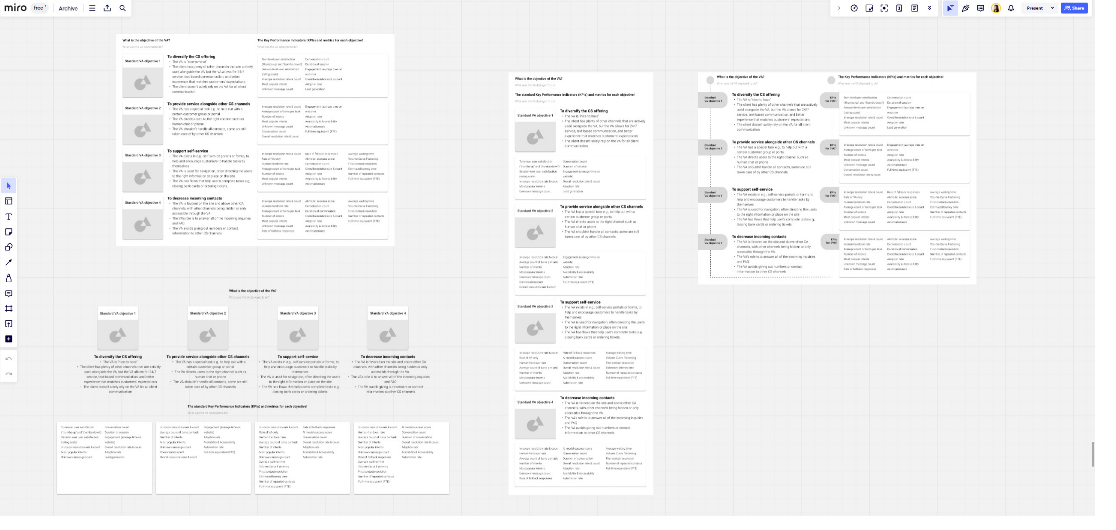

Figure 10 Early iterations of the final prototype displayed in Miro. (Image: Hedvall 2023)

3.5.3 Testing the Final Prototype Solution

The final prototype was tested one more time to gather some ideas on how the prototype could be improved in the short term. The end user's reflections on the produced prototype were gathered in an unstructured manner during a workshop, on CAI objectives and KPIs. Eight AI Trainers from Front AI participated in the workshop and the discussion around the topic, reflecting on the prototype's benefits and areas of development (Figure 11). The purpose of the workshop was to communicate some general information about CAI success, objectives, and KPIs and to get a general understanding of the reception of the prototype among the end users.

During the workshop, the AI Trainers received their own printed handout of the prototype, on which they were invited to write down thoughts and comments. The AI Trainers were also tasked with discussing the topic of maintenance reporting in pairs and using the prototype in the process (Figure 12).

The improvements made to the prototype based on the first round of testing were welcomed and the general feedback on the prototype was positive, with several end users stating that the prototype was logically structured and easy to understand and use. The feedback and findings from the workshop were taken down as notes and added to the list of short-term improvement areas. The final prototype along with the short-term development ideas will be presented in the next section.

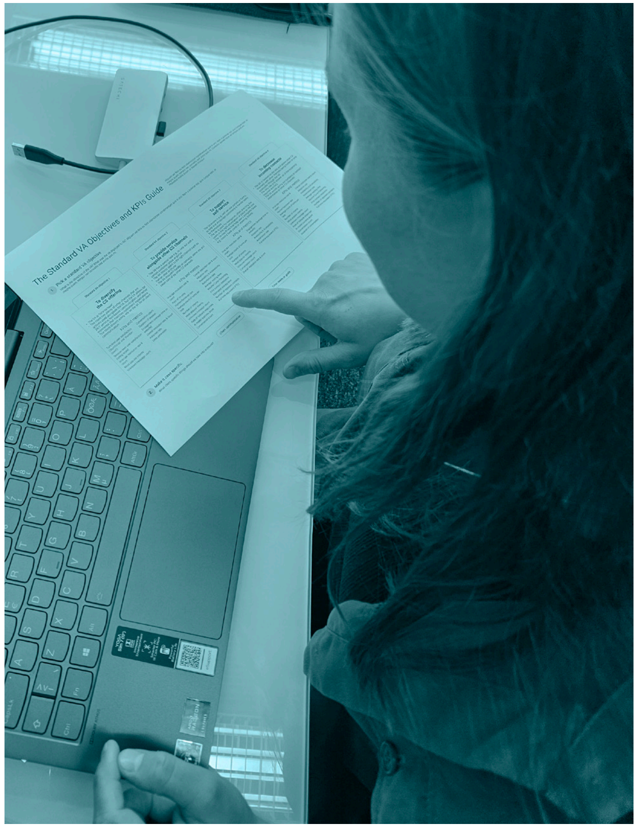

Figure 11 AI Trainers discussing the prototype's benefits and areas of improvement. (Image: Hedvall 2023)

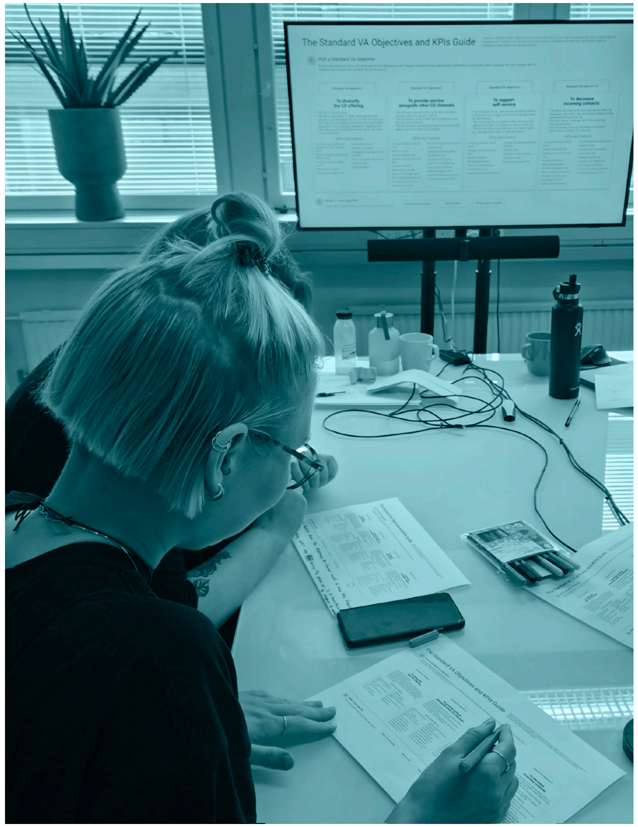

Figure 12 The AI Trainers discussing maintenance reporting in pairs using the prototype. (Image: Hedvall 2023)

3.6 The Result: The Standard CAI Objective and KPI Guide

The final prototype solution, pictured in Figure 13, is a standard CAI objective and KPI guide. The guide's purpose is to help Front AI AI Trainers assess and decide which CAI objectives and adjacent KPIs are relevant for any Front AI client CAI. The prototype introduces a way for Front AI to communicate and increase knowledge around CAI success, objectives, and KPIs.

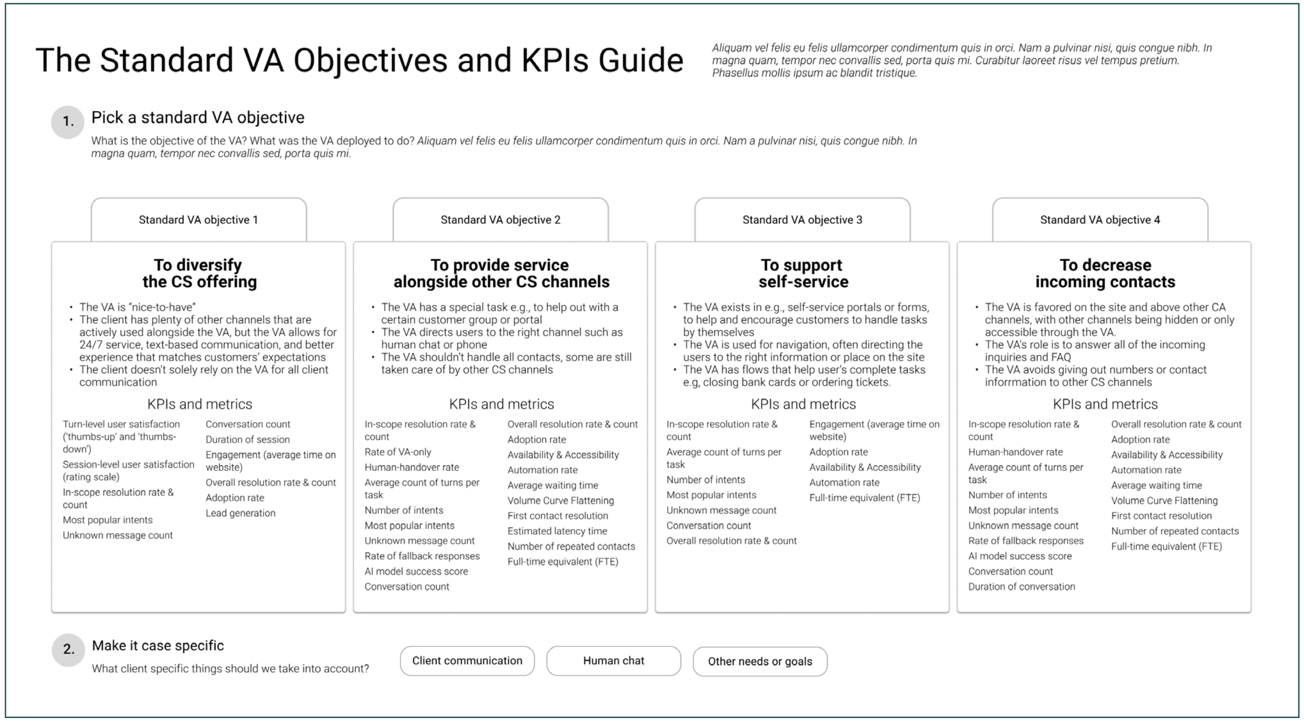

Figure 13 The final prototype: The Standard CAI Objective and KPI Guide. (Image: Hedvall 2023)

3.6.1 Features

The guide presents 4 pre-determined standard objectives. Each objective is rooted in client data collected in Part 1 of this thesis and provides the AI Trainer with an understanding of what the client might wish to achieve with their CAI. The standard objectives are overarching enterprise goals and objectives that determine the success of CAI from the perspective of clients. As these can be seen as the main reasons why clients invest in CAI, it is important for Front AI AI Trainers to understand that these objectives exist and to know how to evaluate their success. The standard objectives help the AI Trainers in the process of picking which objective is most relevant to their client, in cases where no predetermined objectives have been stated.

To help AI Trainers pick the right Key Performance Indicators (KPIs) for the objectives, each standard objective, displayed in the guide, has assigned KPIs that help measure the success of the objective. As enterprise objectives are complex and rarely straightforward, they require a multitude of metrics and KPIs that together form a comprehensive picture of the CAI's performance. By displaying all relevant KPIs for each objective, the AI Trainers are provided with the KPIs without having to think about or search for which KPIs help them evaluate the success of a complex enterprise objective.

The main feature of the prototype is that it uses visual clues to guide the user through the decision-making process. The approach is similar to a workflow often used in user interface (UI) design that guides users from the beginning to the end of a process, completing tasks at each step of the workflow. This approach also helps with discoverability and affordance, as visual clues are used to clearly indicate what the user should do and when. To achieve this, the prototype has numbered call-to-action prompts, accompanied by instructions and further information that guide the user through each step of the workflow.

In addition to the main features – the workflow and the standard objectives – the guide encourages the end user to do some deeper reflection based on their specific case. After the end user has decided on a standard objective for their specific case, they are presented with the option to put everything into context, accounting for any case-specific aspects. The end user is allowed to reflect and make changes to the objectives and KPIS thus further aiding them in the process of gaining a more holistic understanding of their client's CAI. This in turn will help the AI Trainer make a more educated decision on what the client's specific objectives are and what KPIs should be included in measuring the success of those objectives.

The purpose of the guide is to support the end user in the process of deciding on objectives and suitable metrics for any CAI, especially in cases where no predetermined objectives exist. The guide makes the process of choosing and looking for information about CAI objectives and KPIs easier. The guide is a one-pager that can easily be referred to at any moment, allowing the end user to save time and learn in the process.

The guide is tailored to the end user, using language and terms familiar to them. These changes were made based on the findings from the quick prototype testing, that highlighted the importance of making the language understandable to the AI Trainer. The guide also utilizes visual hierarchy to show the importance of key elements, improving interaction.

3.6.2 Limitations and Short-term Development Ideas

While the prototype is a first step in the right direction, it recognizes its limitations.

The guide provides Front AI AI Trainers with the necessary support they need when assessing client objectives and KPIs. However, the prototype does not consider any previous stage in the end user's journey highlighted in the Customer Journey Map. As the guide is heavily focused on helping the end user at the two most crucial steps of the process, it does not put a lot of emphasis on explaining the intricacies of or defining what objectives and KPIs are. Although the user research clearly highlights the need for more sharing of knowledge and information around CAI success, objectives, and KPIs, it is a slightly bigger endeavor and will thus be discussed in depth in the recommendations for future development around the reporting practices at Front AI.

Another area of improvement adjacent to the aforementioned issue is the need for clearer copywriting of the prototype's prompts and content. The prototype should be developed further in the short term by making changes to the content of the guide based on the findings collected during the workshop testing. Any terminology should be clarified e.g., substituting standard VA objectives with Core objectives and any placeholder text should be substituted with user assistance information and content, in accordance with the findings collected during the workshop testing. The content should also be further verified by Front AI and adjusted if necessary. The prototype should also be developed into a fully realized guide, with refined colors and typology.

Based on the feedback collected during the workshop user testing, the development of the guide should pay special attention to the following things: clarifying the purpose of the guide and when it should be used, adding missing case-specific aspects to the second step, and accounting for the before and after steps in the end users' journey of using the guide.

4. CONCLUSIONS

This thesis has delved into the success, evaluation, and impact of CAI chatbots in customer service. Through a comprehensive review of literature and client interviews, the thesis has contributed to the existing body of research and practice-based knowledge around commer-cial CAI evaluation. This paper has highlighted the need for clear evaluation metrics and KPIs to define CAI success with the findings underscoring the importance of a different ap-proach as previously favoured, to evaluate the impact of commercial Conversational AI.

The main findings of this thesis are that literature and practice lack distinction between metrics that contribute to the evaluation of CAI enterprise and business success (the "higher level goals") and KPIs that focus on measuring the CAI's performance on a technological and linguistic level. Firms with commercial CAIs should focus on the enterprise-level KPIs that measure the success of their objectives of diversifying and increasing the efficiency of their customer service, not only the KPIs pertaining to how the CAI performs on its own. To gauge the success of a CAI initiative it needs to be compared to other customer service functions and channels.

The results of this thesis present a standard KPI guide for evaluating CAI success at Front AI. In addition, the results present an understanding of how information should be structured at Front AI to support knowledge acquisition within the AI Trainer team and the de-velopment of reporting practices.

4.1 Suggestions for future development

Future research should continue working toward establishing a clear standard for CAI evaluation. Future research should follow the perspectives approach and focus on creating different standards for different types of CAI chatbots. Creating a standard for commercial CAI evaluation should be emphasized, as it is highly sought after in practice.

The process of increasing understanding of CAI success is a long one. In developing the CAI success and performance reporting practices, Front AI should focus on creating learning opportunities and processes for communicating tacit knowledge around the topic of CAI success and performance. The investigation into the AI Trainers' needs and wants showed that AI Trainers still prefer discussing things together. Providing opportunities to collaborate and to compare findings should be considered in the future development of both the Stand-ard VA objectives and KPIs guide as well as the development of the reporting practices.

The Standard VA objectives and KPIs guide should also be developed in the direction that it can be used in other instances at Front AI. The next step should thus be to investigate how the tool can be made into a company-wide tool and how the findings from this thesis can be utilized in client-facing activities.

5. REFERENCES

Agarwal, S., Agarwal, B. and Gupta, R. 2022. Chatbots and virtual assistants: a bibliometric analysis. Library Hi Tech. Vol. 40 No. 4, 013-1030. Retrieved on 28 July 2023. Available at: https://doi-org.ezproxy.saimia.fi/10.1108/LHT-09-2021-0330

Brill, T. M., Munoz, L., and Miller, R. J. 2019. Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. Journal of Marketing Management. 35:15-16, 1401-1436. Retreived on 28 July 2023. Available at: https://doi.org/10.1080/0267257X.2019.1687571

Butler, C. 2018. What is Design Thinking, and how is it different from service design?. Spotless Says by Medium. Retrieved on 20 April 2023. Available at: https://medium.com/spotless-says/what-is-design-thinking-and-how-is-it-different-from-service-design-e11bc8e3ff5b

Casas, J., Tricot, M-O., Abou Khaled, O., Mugellini, E., Cudre-Mauroux, P. 2020. Trends & Methods in Chatbot Evaluation. Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI '20 Companion, Netherlands, October 25–29, 280-286, 2020. Retrieved on 28 July 2023. Available at: https://doi.org/10.1145/3395035.3425319

Clatworthy, S. 2013. Design support at the front end of the New Service Development (NSD) process: the role of touch-points and service personality in supporting team work and innovation processes. Doctoral thesis. The Oslo School of Architecture and Design. Oslo. Retrieved on 20 April 2023. Available at: https://aho.brage.unit.no/aho-xmlui/handle/11250/93069

Dam, R. F. 2022. 5 stages in the design thinking process. The Interaction Design Foundation. Retrieved on 2 September 2023. Available at: https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process

Dam, R.F., Siang, T.Y. 2019. Stage 2 in the Design Thinking Process: Define the Problem and Interpret the Results. The Interaction Design Foundation. Retrieved on 25 August 2023. Available at: https://www.interaction-design.org/literature/article/stage-2-in-the-design-thinking-process-define-the-problem-and-interpret-the-results

De Andrade, I.M. and Tumelero, C. 2022. Increasing customer service efficiency through artificial intelligence chatbot. Revista de Gestão, Vol. 29, No. 3, 238–251. Retreived on 20 April 2023. Available at: https://doi.org/10.1108/REGE-07-2021-0120

Driver, A. 2018. Design Drivers: What drives great design?. Team Consulting. Retrieved on 10 September 2023. Available at: https://www.team-consulting.com/insights/design-drivers-what-drives-great-design/

Elmansy, R. 2023. Using the Affinity Diagram to Organize Ideas. Designorate. Retrieved on 28 July 2023. Available at: https://www.designorate.com/using-the-affinity-diagram/

Ferreira, B., Silva, W., Oliveira, E., Conte, T., 2015. Designing Personas with Empathy Map. Conference paper. SEKE 2015.: 27th International Conference on Software Engineering and Knowledge Engineering. Pittsburg, USA, Jul 6-8, 2015. Retrieved on 25 August 2023. Available at: https://www.researchgate.net/publication/276207468_Designing_Personas_with_Empathy_Map

Følstad, A., Skjuve, M., Brandtzaeg, P.B. 2019. Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Under-stand Interaction Design. International Conference on Internet Science, INSCI 2018: Internet Science, Lecture Notes in Computer Science, vol 11551. 145-156. Retrieved on 13 April 2023. Available at: https://doi.org/10.1007/978-3-030-17705-8_13

Gartner. 2023. 2021 Customer Service Trends and Priorities. Gartner. Retrieved on 13 April 2023. Available at: https://www.gartner.com/en/customer-service-support/insights/service-leaders-priorities

Gioia, D. A., Corley, K. G., Hamilton, A. L. 2013. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. Organizational Research Methods, 16(1), 15-31. Retrieved on 30 April 2023. Available at: https://doi.org/10.1177/1094428112452151

Gray, D. 2017. Updated Empathy Map Canvas. Medium. Retrieved on 25 August 2023. Available at: https://medium.com/@davegray/updated-empathy-map-canvas-46df22df3c8a.

Hamel, J., Dufour, S. and Fortin, D. 1993. Case Study Methods. Qualitative Research Methods. Newbury Park, CA: SAGE Publications, Inc. Retrieved on 28 July 2023. Available at: https://doi.org/10.4135/9781412983587

Hung, V., Elvir, M., Gonzalez, A., DeMara, R. 2009. Towards a Method For Evaluating Naturalness in Conversational Dialog Systems. 2009 IEEE International Conference on Systems, Man and Cybernetics. San Antonio, TX, USA, October 11-14, 2009, 1236-1241, Retrieved on 13 April 2023. Available at: https://doi.org/10.1109/ICSMC.2009.5345904

Interaction Design Foundation. 2019. What are Customer Journey Maps? The Interaction Design Foundation. Retrieved on 10 September 2023. Available at: https://www.interaction-design.org/literature/topics/customer-journey-map

Jadeja, M., Varia, N. 2017. Perspectives for Evaluating Conversational AI. ICTIR' 17 Workshop on Search-Oriented Conversational AI (SCAI' 2017). Amsterdam, Netherlands. Retrieved on 28 July 2023. Available at: https://www.researchgate.net/publication/319736219_Perspectives_for_Evaluating_Conversational_AI

Janssen, A., Grützner, L. and Breitner, M. 2021. Why do Chatbots fail? A Critical Success Factors Analysis. International Conference on Information Systems (ICIS). Austin, TX, US, December 12-15, 2021. Retrieved on 13 April 2023. Available at: https://www.researchgate.net/publication/354811221_Why_do_Chatbots_fail_A_Critical_Success_Factors_Analysis

Kolari, J. and Kallio, A. 2023. Tekoäly 123 — Matkaopas tulevaisuuteen. First Edition. Jyväskylä: Docendo.

Kisling, D. 2022. Types of Bots for Business: AI Chatbot or Button-Based One?. Expert opinion blog post. Tovie AI. Retrieved on 12 April 2023. Available at: https://tovie.ai/blog/types-of-chatbots-for-business

Kurronen, J., Erkkilä, K., Moisio-Imbert, K. 2015. Palvelumuotoilun työkalupakki. peda.net. Retrieved on 25 August 2023. Available at: https://peda.net/id/b8b31b72a4d

Lahden ammattikorkeakoulu, ProAgria Etelä-Suomi ry. 2018. PALMA -Palvelumuotoilulla uutta liiketoimintaa maaseudulle: Palvelumuotoilun Sanasto. www.palma.fi. Retreived on 28 August 2023. Available at: https://www.palma.fi/palma-malli/kayttajalahtoisyys/Palvelumuotoilun%20sanasto.pdf

Lewandowski, T., Heuer, M., Vogel, Pascal, and Böhmann, T. 2022. Design Knowledge for the Lifecycle Management of Conversational Agents. The 17th International Conference on Wirtschaftsinformatik. Nürnberg, Germany, February 2022. 3. Retrieved on 13 April 2023. Available at: https://aisel.aisnet.org/wi2022/ai/ai/3

Lewandowski, T., Poser, M., Kučević, E., Heuer, M., Hellmich, J., Raykhlin, M., Blum, S., Böhmann, T. 2023. Leveraging the Potential of Conversational Agents: Quality Criteria for the Continuous Evaluation and Improvement. The 56th Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, United States, January, 2023, 3444-3453. Available at: https://hdl.handle.net/10125/103055

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., Theeramunkong, T. 2019. A Survey on Evaluation Methods for Chatbots. Proceedings of 2019 7th International Conference on Information and Education Technology. ICIET 2019. Aizu-Wakamatsu, Japan March 29-31, 2019. 111-119. Retrieved on 13 April 2023. Available at: https://doi.org/10.1145/3323771.3323824

McKinsey & Company. 2022. The state of AI in 2022—and a half decade in review. McKinsey & Company. Retrieved on 12 April 2023. Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/

Meyer von Wolff, R., Hobert, S., Schumann, M. 2021. Sorry, I Can't Understand You! – Influencing Factors and Challenges of Chatbots at Digital Workplaces. Ahlemann, F., Schütte, R., Stieglitz, S. (eds) Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation, vol 47. Springer, Cham. Retrieved on 13 April 2023. Available at: https://doi.org/10.1007/978-3-030-86797-3_11

Miaskiewicz, T. and Kozar, K.A. 2011. Personas and user-centered design: How can personas benefit product design processes?. Design studies, 32(5), 417-430. Retrieved on 25 August 2023. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0142694X11000275

Narges, I., Ciper, E., Walker Wilson, K., Klein, D. 2018. Service Design Tools & Methods. Image. Retrieved on 20 April 2023. Available at: https://medium.com/capitalonedesign/service-design-tools-methods-6e7f62fcf881

Osei-Mensah, B., Asiamah, E.O. and Sackey, R., 2023. Strategic Communication and Artificial Intelligence: Reviewing Emerging Innovations and Future Directions. Archives of Business Research, 11(1), 85-102. Retrieved on 13 April 2023. Available at: https://doi.org/10.14738/abr.111.13616

Peras, D. 2018. Chatbot Evaluation Metrics. Economic and Social Development: Book of Proceedings. 36th International Scientific Conference on Economic and Social Development - "Building Resilient Society". Zagreb, December 14-15, 2018. 89-97. Retrieved on 13 April 2023. Available at: https://d1wqtxts1xzle7.cloudfront.net/58010665/Book_of_Proceedings_esdZagreb2018_Online-libre.pdf?1545035913=&response-content-disposition=inline%3B+filename%3DEXAMINING_THE_EFFECTS_OF_TRAVEL_CONSTRAI.pdf&Expires=1698152893&Signature=XomflViV7Y9H5GnnA6fCrPaTlzlpO5Kq9e4FVK8iykkqRAfq5vqaWX28e1lUACKPcTskYLXQzzFqcz-qHtpWzsmrkfkSd~BPBte3s2kGbscDe-mEN0JUPq-IyxswROj80DCT2m1m6DbGwpkqXCGVzRgwkEmMRpNZ0gEeQE5sIl~T85tKIwBR32ZgRtO8hf84sCQrFSki2DCiFqOVa8EtoGjXc-1cKztNUWpKcGIFDkKKznvfdP5FqmP6vHjTmb2ailTvVlbhi6ExTB0rrTNMKjfrIS9PRHBkA~Xm7SrBa38cnHv5x8zcM2~pSodYqzjZ6iq0HPBtYMH0x2Hqx8FGtA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Punch, K. F. 2013. Introduction to social research: Quantitative and qualitative approaches. Sage, London.

Radziwill, N., Benton, M. 2017. Evaluating Quality of Chatbots and Intelligent Conversational Agents. Retrieved on 20 April 2023. Available at: https://www.researchgate.net/publication/316184347_Evaluating_Quality_of_Chatbots_and_Intelligent_Conversational_Agents

Russell-Rose, T. 2017. A framework for chatbot evaluation. WordPress.com. Retrieved on 31 July 2023. Available at https://isquared.wordpress.com/2017/01/24/a-framework-for-chatbot-evaluation/

Saunders, M., N. K., Lewis, P., Thornhill, A., 2015. Research Methods for Business Students. eBook. 7th edition. Pearson Education. ISBN1292016647

Schütte, R., Stieglitz, S. 2021. Innovation Through Information Systems. Lecture Notes in Information Systems and Organisation, vol 47. Springer, Cham. Retreived on 13 April 2023. Available at: https://doi.org/10.1007/978-3-030-86797-3_11

SDT. 2023a. Synthesis Wall | Service Design Tools. Servicedesigntools.org. Retrieved on 28 July 2023. Available at: https://servicedesigntools.org/tools/synthesis-wall

SDT. 2023b. Journey Map | Service Design Tools. Servicedesigntools.org. etrieved on 10 September 2023. Available at: https://servicedesigntools.org/tools/journey-map

SDT. 2023c. Mindmap | Service Design Tools. Servicedesigntools.org. Retrieved on 10 September 2023. Available at: https://servicedesigntools.org/tools/mindmap

Sherman, N. (reporter) 2023. Google's Bard AI bot mistake wipes $100bn off shares. BBC News. Online news article 8 Feb. Retrieved on 31 July 2023. Available at: https://www.bbc.com/news/business-64576225

SocialUP. 2023a. Samankaltaisuuskaavio. Designthinking-socialup.eu. Retrieved on 28 July 2023. Available at: http://toolkit.designthinking-socialup.eu/fi/saman-kaltaisuus-kaavio

SocialUP. 2023b. Käyttäjäpersoonat. Designthinking-socialup.eu. Retrieved on 28 July 2023. Available at: http://toolkit.designthinking-socialup.eu/fi/k%C3%A4ytt%C3%A4j%C3%A4-persoonat

Stevens, E. 2019. How To Define A Problem Statement: Your Guide To The Second Step In The Design Thinking Process. careerfoundry.com. Retrieved on 25 August 2023. Available at: https://careerfoundry.com/en/blog/ux-design/stage-two-design-thinking-define-the-problem/

Stickdorn, M., Hormess, M.E., Lawrence, A. and Schneider, J. 2018. This is service design doing: applying service design thinking in the real world. O'Reilly Media, Inc. ISBN1491927186

Stickdorn, M. and Schneider, J. 2012. This is service design thinking: Basics, tools, cases. John Wiley & Sons. ISBN9781118156308

Suhaili, S.M., Salim, N. and Jambli, M.N. 2021. Service chatbots: A systematic review. Expert Systems with Applications. Vol. 184, 115461. Retrieved on 28 July 2023. Available at: https://doi.org/10.1016/j.eswa.2021.115461

Vilser, M., Fuchs, S., Wittges, H., Krcmar H. 2022. Evaluation of IBM's Conversational AI Framework. Master's Thesis in Information Systems. Technical University of Munich, Department of Informatics. Munich, Germany. Retrieved on 13 April 2023. Available at: http://doi.org/10.13140/RG.2.2.11813.99040

5.1 Tables

Table 1, Peras, D., modified by Hedvall A. 2018. The Business perspective in the The Chatbot Evaluation Framework by Peras (2018)

Table 2, Hedvall, A., 2023. An overview of the case firms and interviewees selected for the client interviews.

Table 3, Hedvall A., 2023. CAI objectives identified in interviews with Front AI clients.

Table 4, Hedvall A., 2023. CAI performance KPIs and metrics.

Table 5, Hedvall A., 2023. KPIs that measure CAI success.

Table 6, Hedvall A., 2023. KPIs and metrics for Front AI for evaluating CAI success and performance.

5.2 Figures

Figure 1, Hedvall, A., 2023. Perspectives and categories for CAI evaluation at Front AI based on Peras' (2018) Chatbot evaluation framework (CEF)

Figure 2, Interaction Design Foundation edited by Hedvall, A., 2022. The design thinking process developed by Hasso Plattner Institute of Design at Stanford (d.school). Retrieved on 23 October 2023. Available at: https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process

Figure 3, Gray, D., modified by Hedvall, A., 2017. The Empathy Map Canvas by Gray (2017) filled in with the findings from the contextual interviews. Retrieved on 23 October 2023. Available at: https://cdn-images-1.medium.com/v2/resize:fit:1600/1*I1ffOWdPWQva3dCMQE-TAQ.png

Figure 4, Hedvall, A., 2023. The Persona: Jessie.

Figure 5, Hedvall, A., 2023. The design drivers for the prototype.

Figure 6, Hedvall, A., 2023. A snapshot of The Customer Journey Map with ideas and pain points highlighted in red.

Figure 7, Hedvall, A., 2023. The mindmap with the ideas marked as either approved, rejected, out-of-scope or other.

Figure 8, Hedvall, A., 2023. AI Trainer exploring the quick prototype in Miro.

Figure 9, Hedvall, A., 2023. The AI Trainers are giving feedback on the look and feel of the prototype.

Figure 10, Hedvall, A., 2023. Early iterations of the final prototype displayed in Miro.

Figure 11, Hedvall, A., 2023. AI Trainers discussing the prototype's benefits and areas of improvement.

Figure 12, Hedvall, A., 2023. The AI Trainers discussing maintenance reporting in pairs using the prototype.

Figure 13, Hedvall, A., 2023. The final prototype: The Standard CAI Objective and KPI Guide.

5.3 Appendices

Appendix 1, Peras, D., 2018. The Chatbot Evaluation Framework by Peras (2018). Chatbot Evaluation Metrics. Economic and Social Development. Book of Proceedings. 36th International Scientific Conference on Economic and Social Development. Zagreb, 14-15 December 2018.

6. APPENDICES

Appendix 1 The Chatbot Evaluation Framework by Peras (2018)

Perspective w	Category	Attributes	Metrics	Approach
User experience perspective	Usability	• task completion • getting assistance or information • support of a minimal set of commands • response type frequency	• response type relative frequencies • percentage of match • response type relative probability • rating scale • surveys • questionnaires • support of Help and Cancel commands	Qualitative, Quantitative
	Performance	• robustness • responses in unexpected situations • coherence • effective task allocation	• percentage of success • rating scale	Qualitative
	Affect	• personality • emotional information • entertainment • engagement • personality traits • human assistance provision • trustworthiness	• rating scale • surveys • questionnaires • checking for keywords · number of dialogue turns • total conversation duration	Qualitative, Quantitative
	Satisfaction	• expectation • impression • command • navigability • engagement • entertainment • curiosity • social relations • ability to learn • ability to aid	• conversation duration • number of conversation turns • rating scale	Qualitative, Quantitative
Information retrieval perspective	Accuracy	• ability to foresee language variations	• precision • recall • typing errors and synonyms	Quantitative
	Accessibility	• ability to detect meaning and intent and to respond appropriately	• context sensitiveness • percentage of success • number of inappropriate responses • turn correction ratio	Quantitative
	Efficiency	• how well the resources are applied to achieve the goals	• matching types • measuring the answer time for the commands and obtaining mean values • total elapsed time • total number of users turns • total number of turns per task	Quantitative
Linguistic perspective	Quality	• correctness of the responses • categorization of responses	• Likert scale	Qualitative
	Quantity	• adequateness of information	• Likert scale	Qualitative
	Relation	• relevancy of responses to the context of the conversation	• Likert scale	Qualitative
	Manner	• unambiguity of the responses	• Likert scale	Qualitative
	Grammatical accuracy	• acceptability from grammatical and meaning perspective	• total number of errors made in the chat period • Word-level analysis (vocabulary range, spelling, upper/lower case) • Grammar-level analysis (nouns, pronouns, verbs, word order, etc.)	Qualitative
Technology perspective	Humanity	• naturalness • maintaining themed discussion • responding to specific questions • non-understanding rate	• Turing Test • rating scale • percentage of success • percentage of rejection	Qualitative, Quantitative
Business perspective	Business value	• efficiency cost • qualitative cost	• number of users • duration of the chatbot conversation • number of the chatbot conversations • number of the agents included in conversation • duration of the conversation with an agent • number of the unsuccessful conversations • number of the unsuitable responses • number of repeated queries	Quantitative