SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Informatics

# Intelligent Channel Navigation in Customer Service Using Large Language Models

**Constantin Ehmanns**

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Informatics

# Intelligent Channel Navigation in Customer Service Using Large Language Models

# Intelligente Kanalnavigation im Kundenservice mit Großen Sprachmodellen

| | |
|---|---|
| Author: | Constantin Ehmanns |
| Supervisor: | Prof. Dr. Florian Matthes |
| Advisor: | Nektarios Machner |
| Submission Date: | 28.02.2025 |

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Salt, 28.2.2025

Location, Submission Date

Author

# AI Assistant Usage Disclosure

## Introduction

Performing work or conducting research at the Chair of Software Engineering for Business Information Systems (sebis) at TUM often entails dynamic and multi-faceted tasks. At sebis, we promote the responsible use of *AI Assistants* in the effective and efficient completion of such work. However, in the spirit of ethical and transparent research, we require all student researchers working with sebis to disclose their usage of such assistants.

For examples of correct and incorrect AI Assistant usage, please refer to the original, unabridged version of this form, located at this link.

## Use of *AI Assistants* for Research Purposes

**I have used AI Assistant(s) for the purposes of my research as part of this thesis.**

x Yes No

**Explanation:**

- DeepL: Translation of words and sentences.
- ChatGPT: Improvements and modifications of (parts of) prompts. Each usage is detailed in the thesis.

I confirm in signing below, that I have reported all usage of AI Assistants for my research, and that the report is truthful and complete.

Saal, 28.2.2025
Location, Date

C. Eh
Author

# Acknowledgments

I would like to express my gratitude to Prof. Matthes who made this interesting project possible and enabling my master's thesis in the first place. I also wish to extend my special thanks to my advisor Nektarios Machner who guided me throughout the entire project and continuously provided valuable feedback and support.

I would also like to thank our contact persons at the case study company for motivating and guiding the thesis from the enterprise side while providing valuable insights and constructive feedback along the process.

Eventually, I would like to thank my family and friends for the extensive support I received over the past couple of years.

# Abstract

This thesis touched upon three topics in the area of customer service in collaboration with a large European insurance company: channel decision factors, Large Language Model (LLM)-powered channel preference prediction and generation of reasons for such preferences. A structured literature review was conducted to uncover potentially relevant customer choice determinants for channel selection and combined with a workshop and interviews in the case study company. An LLM-based communication channel classifier for incoming requests in customer service was developed and various prompt and input strategies tested for appropriateness. The inclusion of many user related data points, examples (few-shots) and Chain-of-Thought prompting proved to be a strong combination and delivered on-par performance with simple learning-based approaches such as regression, albeit both scored moderate results in general. The LLM was further tested for its capability of generating similar reasons to those provided by real humans which worked well for simple reasons but seems to be challenging overall.

# Kurzfassung

In dieser Arbeit wurden in Kollaboration mit einem großen europäischen Versicherungsunternehmen drei Themen im Bereich Kundenservice behandelt: Faktoren für Kanalentscheidungen, Vorhersage von Kanalpräferenzen mit Hilfe von Großen Sprachmodellen (Large Language Models, LLMs) und Generierung von Gründen für solche Präferenzen. Eine strukturierte Literaturrecherche wurde durchgeführt, um potenziell relevante Entscheidungsfaktoren für die Kanalwahl zu ermitteln, und mit einem Workshop und Interviews in dem Fallstudienunternehmen kombiniert. Es wurde ein LLM-basierter Kommunikationskanal-Klassifikator für eingehende Anfragen im Kundenservice entwickelt und verschiedene Prompt- und Eingabestrategien auf ihre Nützlichkeit getestet. Die Einbeziehung von möglichst vielen Nutzerdaten, Beispielen (few-shots) und Chain-of-Thought-Prompting erwies sich als eine aussichtsreiche Kombination und lieferte eine vergleichbare Leistung wie einfache lernbasierte Ansätze wie z.B. Regression, obwohl beide im Allgemeinen eher mäßige Ergebnisse erzielten. Das LLM wurde außerdem auf seine Fähigkeit hin getestet, ähnliche Gründe zu generieren, wie sie von echten Menschen angegeben werden, was bei einfachen Gründen gut funktionierte, aber insgesamt eine Herausforderung darzustellen scheint.

# Contents

# 1. Introduction

## 1.1. Problem

Steve Jobs famously said: "Get closer than ever to your customers. So close that you tell them what they need well before they realize it themselves" [1]. In the world of customer service, knowing what the customer needs goes beyond answering a question or service fulfillment. It can also mean the selection of the appropriate communication channel to address the concerns in. For service providers such as insurance companies or banks, customer service often provides the interface towards its customers and thus has a great influence on the reputation and perception of the company [2, 3]. Hence, the quality of the provided service has real business impact which can be assessed by for example looking at the time it takes for the company to answer the customer's request and the level of satisfaction customers experience when receiving assistance [4]. While companies provide an array of communication channels such as the traditional (call) hotline, chat services or self-service opportunities via websites and apps, the most natural one is the hotline [4]. For companies with high quality service standards and majority human agents handling the incoming phone calls, this poses a series of problems: significant workforce resources, potentially long waiting times or bottlenecks in highly frequented periods and an under-usage of other channels which have been carefully designed to handle certain workflows as efficient as possible. To this end, we are collaborating with a large European insurance company to address the problem of finding the right channels for each customer individually. The case study company has not yet implemented any system to automatically route customers to different channels in case it deems this appropriate (or suggest the option) nor was any structured data available to aid in the development of such a system.

## 1.2. Objectives

In the context of the case study company, we aim to answer three questions:

- In customer service centers, what are relevant factors for deciding the optimal channel for customer service requests?
- How do different input factors and prompt strategies influence the effectiveness of LLMs in selecting appropriate communication channels for customer service requests?
- How well do LLMs predict the reasons for choosing a customer service channel?

Analyzing determinants of customer channel choice will allow for a better understanding of the research area and provide a basis for the type of information that will need to be collected to create an exhaustive dataset. This will then need to be annotated and labeled with real customer data which is the goal of a data collection process. The implemented Large Language Model (LLM)-based channel suggestion agent is designed to analyze a situation in which a customer is calling the (phone) hotline and predicts whether other channels may be appropriate for issue resolution. This opens interesting further investigations into the usability of LLMs in customer service processes such as predicting the actual reasoning of a customer behind a channel choice. The indicated solution also presents a first step into a seamless (omni-)channel switching experience.

## 1.3. Outline

Chapter 1 motivates the thesis and its objectives. The *Theoretical Foundation* in Chapter 2 describes important concepts for the field of customer service and insights into the emergence and capabilities of Large Language Models to lay a basis and common understanding of the examined topics. This is followed by Chapter 3 - an overview of the *Related Work*. In Chapter 4, the approach to the problem is described as design science research inspired and explains the mode of collaboration with the case study company. The synthesis of the customers decision factors in channel choice is described in detail (Chapter 5) as well as the development of the LLM-based channel concierge and reason prediction agent (Chapter 6). In Chapter 7, the process of how the data was obtained is described along with a characterization of the data set. In Chapter 8, the LLM agent is then evaluated based on the data to show the usability of the proposed solution and investigate the appropriateness of LLMs in this field of application. In Chapter 9, the implications of this work and future research directions are explored and Chapter 10 draws a conclusion.

# 2. Theoretical Foundation

This thesis touches upon two major streams of research: customer service/care[1] and Large Language Models (LLMs). Customer care constitutes the domain of the problem whereas LLMs are providing the solution space. Understanding the history and state of the art in both areas is key to tap into the possibilities and challenges that present itself in the customer care channel choice space. The following sections will provide an overview of the relevant developments in both customer care and LLMs.

## 2.1. Customer Care

Customer care entails a company's provisioning of information, tools and services to its customers [5]. In its realm, the quality of the provided services is highly important because of the effects it has on customer retention or financial performance [6]. The methods used for the communication between company and customers are called *channels* and can take the form of physical stores, websites, chat and phone support, online communities and more [7]. Companies, especially retailers, have reacted to the upcoming of the online channel by adopting multi-channel approaches, evolving from mere addition of new channels to the channel portfolio towards cross-channel customer management [8]. This is succeeded by the move towards omni-channel strategies which seeks to provide a cohesive cross-channel experience to the user [9, 10]. Gerea et al. [11] additionally argue that the adoption of the omnichannel approach through established companies was motivated by the intent to keep pace with new and very successful digital players (the likes of Google and Amazon) and improve their customer service [11].

Multi- and Omnichannel research is a broad field with a variety of different focus points. Topics range from analyzing integration quality and switching behavior in omni- and multi-channel environments [12, 9, 13] to customer channel behavior in the context of buy-and-return [14, 15] or complaining [16]. A significant amount of effort has also been made to understand channel choice determinants from a customer point of view although a comprehensive collection of the involved factors has yet been only addressed by Wolf and Steul-Fischer [17].

On top of the changes in the number of available channels and their integration, technological advances have enabled numerous innovations in the channels itself [18]. Improvements in technology have led to increased personalization and thus better service quality in general [18]. Conversation channels such as chat and phone calls between the customer and the

[1] Customer care and service are used mostly synonymously throughout the thesis

company constitute an especially interesting field of advances. Chat services have become a very popular method of communication for the real-time provision of information or assistance. While these can be human operated, efforts have been made to use software instead. [19] Weizenbaum [20] already built a rudimentary chatbot in the 1960s which employed a rule-based conversation system hinging on keywords appearances [20]. From there, chatbots continuously improved with technological advances and have evolved into digital assistants with increasingly more natural interactions [19, 21, 22, 23]. Today´s chatbots equipped with AI technology can automatically handle a variety of customer requests with a significantly reduced workforce [24]. Similarly to chat services, call centers have evolved as well. Traditionally, these were operated with an interactive voice response (IVR) system that enables users to navigate to its services by the means of the phone keypad with the fallback option being a routing to a human call agent. With the adoption of AI systems, users can now engage in human-like conversations to obtain the required services through Natural Language Processing. [25]

## 2.2. Large Language Models

Natural Language Processing (NLP) is a field of reseach in the general realm of Artificial Intelligence (AI) which is concerned with both the analysis and manipulation of human language or speech [26, 27]. Two important subfields are Natural Language Understanding (NLU) and Natural Language Generation (NLG), researching the automated meaning extraction of language and the process of generating natural language respectively [28]. Language modelling (LM) in particular "aims to model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens" [29]. These models are of special interest since developments in Large Language Models (LLMs) and large Pre-trained transformer-based Language Model (PLMs) specifically are heavily influencing the entire field of NLP [30, 27]. Karanikolas and Tousidou even argue that LLMs are closely entangled with the foreground of research in AI in general [27].

### 2.2.1. Recent History

To contextualize recent advances and accomplishments in this field, understanding the processes that lead to them are crucial. Zhao et al. [29] identified four stages of development in the field of language modelling research: Statistical Language Models (SLM), Neural Language Models (NLM), Pre-trained Language Models (PLM) and Large Language Models (LLM) [29]. *Statistical Language Models* leverage the Markov assumption to predict the next word dependant on the previous context. This is called an n-gram language model in case a context of length n is considered [29]. The Markov assumption states, that the probability of a next word is (assumed to be) dependant only on the one before [31]. The probability of a word can thus be approximated as follows (N being the n-gram size) [31]:

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-N+1:n-1})$$

Borrowing an example from Jurafsky and Martin [31], we can approximate the probability $P(\text{blue} \mid \text{The water of Walden Pond is so beautifully})$ with the probability $P(\text{blue} \mid \text{beautifully})$. *Neural Language Models* make use of neural networks to describe the likelihood of word sequences [29]. Bengio et al. [32] for example address some shortcomings of the n-gram approach: not accounting for words outside of (before) their limited context and ignoring similarity between words. They try to overcome these by representing words as real-valued feature vectors with a fixed and finite dimension. The probability distributions used for deciding on the next word are calculated using neural networks that process the feature vectors of the previous words (context) [32]. *Pre-trained Language Models* build on the concept of transfer learning in which knowledge gained from one problem can be leveraged in new settings [33]. In language modelling, this translates to training generative models on large text corpora and applying the resulting representations to problems where there are fewer labeled data [34]. Prominent examples such as the Generative Pre-trained Transformer (GPT) [35] combine unsupervised pre-training on large unlabeled text-corpora with supervised fine-tuning for the target tasks [35]. GPT models employ the Transformer architecture [36] which has improved treatment of long-term dependencies over previous approaches such as recurrent networks [35]. Scaling PLMs yields impressive results which has led to the creation of the term *Large Language Model* [29]. Kaplan et al. [37] show that scaling of the following has the greatest effect on the performance of the (Transformer architecture) models: "the number of model parameters $N$ (excluding embeddings), the size of the dataset $D$, and the amount of compute $C$ used for training" [37]. The performance boosts hold as long as both $N$ and $D$ are increased in parallel (though not necessarily with the same scaling factor) [37]. This brought about models such as GPT-3 [38] that features 175 billion parameters and was trained on several datasets with cumulative 300 billion tokens or PaLM with 540 billion parameters and 780 billion of training tokens [38, 39]

### 2.2.2. Use Cases

Large Language Models prove to be very versatile. They are capable – in varying degrees - of "language generation, knowlegde utilization,[...] complex reasoning" [29], "human alignment, interaction with the external environment, and tool manipulation" [29]. Language generation includes language modeling (finding the next tokens), conditional text generation (such as translation, summarization or question answering) and code synthesis (formal computer programs). Knowledge utilization tasks may relate to question answering in open- and closed-book settings (i.e. the LLM having access to external resources such as documents or not) as well as knowlegde completion which pertains to filling-in incomplete knowlegde artefacts. Complex reasoning in the realm of LLMs (may it be knowlegde, symbolic or mathematical) allows for making decisions based on some logic or facts and the ability of human alignment comprises the conformity to values and needs. Through interactions with the external environment (s.a. feedback) and tool manipulations like API-calls, LLMs increase their area of influence.[29] This makes them useful in a variety of application areas. Code generation tools like GitHub CoPilot [40] are able to provide real-time code and comment suggestions for source code creation. GitHub CoPilot is powered by Codex, a GPT model

fine-tuned on code [41]. Systems like TransLlama [42] leverage (minimally) fine-tuned LLMs to achieve impressive language translation results. Bing[2] uses an LLM to present a comprehensive view of search results [43].

### 2.2.3. Limitations

Despite the promising technology and broad range of applicability, LLMs face a series of limitations. Hadi et al. [44] compiled a list of noteworthy drawbacks. We would like to point out a selection of these deemed most relevant to the application at hand: bias, resource needs, explainability and hallucinations. Biases existent in training data can lead to their presence in the corresponding machine learning algorithms [45]. Caliskan et al. [46] demonstrated the existence of a series of linguistic biases in a word embedding model. Amongst their findings were the implicit association of European American names with pleasant words and those of African American names with unpleasant ones. Also, they uncovered gender biases such as the association of female names with family words and male names with corresponding career words. [46] Additionally, the vast resource needs of LLMs are not to be underestimated. The energy used in training the BLOOM [47] LLM is reported to have ben around 433000 kWh[3] and resulted in the emission of close to 25 tonnes of $CO_2$ equivalents. During deployment, not only the actual processing of requests but also the idle time (no requests) has a significant energy consumption. [49] In terms of explainability, LLMs pose a challenge because of the high model complexity and in-transparent underlying processes [50]. Doshi-Velez and Kim [51] define interpretability[4] in the context of machine learning "as the ability to explain or to present in understandable terms to a human" [51]. Compared to even regular deep learning algorithms, the sheer size of LLMs (both in number of parameters and amount of training data) complicates the understanding of the concrete processes that lead to the results [50]. Finally, hallucinations pose another threat to the applicabiltiy of LLMs. Hallucinations in NLG are traditionally categorized into intrinsic and extrinsic. Instrinsic hallucinations alter information from the input while extrinsic halluconations refer to the inclusion of information that is not present in the input [52, 53]. Huang et al. [53] further propose a redefined categorization for hallucinations to address more LLM-specific challenges: factuality and faithfulness. Factuality hallucinations describe outputs which are contradicting exsiting facts or cannot be verified. On the other hand, faithfulness hallucinations refers to the deviation of the LLM output from the user's intent or inconsitencies within the (LLM) answer. [53]

[2]Search engine from Microsoft; https://www.microsoft.com/en-us/edge/features/the-new-bing
[3]For reference: in 2023, the per-person power consumption in Germany was approximatley 6100 kWh [48]
[4]Consider synonymous to explainability, as in [50]

# 3. Related Work

Literature on automated channel switching or suggestion is – to the best of the author's knowledge – sparse and there is no record of Large Language Model usage for classification in the channel navigation space. Nevertheless, either space – LLMs for classification and automated channel navigation – has been researched before individually.

## 3.1. LLM Classification

Despite the generative nature of language models (with respect to the output), LLMs have been applied to various classification tasks such as intent classification [54, 55], sentiment analysis [56, 57] or product categorization [58]. In health care for example, Rao et al. [59] have tested the viability of LLMs in clinical decision support to select the appropriate radiology service in the area of of breast pain and breast cancer screening. They provided OpenAIs ChatGPT with some patient data (such as the patients lifetime risk of cancer) and asked for a suitable radiology service for this situation [59]. In finance, Fatemi et al. [60] further investigate how (different modified and fine-tuned) LLMs can proof useful on a variety of financial text classification problems such as hawkish-dovish or deal completion classification [60].

## 3.2. Automated Channel Navigation

The idea of using NLP methods (such as automatic speech recognition and machine-learning classifiers) to route calling users to the correct human agent or alternatively self-service options is not new as Tyson and Matula [61] have already demonstrated in 2004 with their Latent Semantic Indexing approach. They do not, however, open the possibility to switch the communication channel [61]. More recently, Liu et al. [4] leveraged deep reinforcement-learning to create channel recommendations for calling customers at a financial technology company based on information about the available channels (estimation of the incoming request volume) and the user (age, gender, province, household car, assets and credit limit). The proposed solution outperforms the baseline mechanism which is a rule-based system built by the business in terms of number of people accepting a channel switch suggestion. The main goal here is to avoid congestion in bottleneck channels. The authors suggest potential improvements to their approach by - amongst others - additionally considering the actual user request in text form on top of the previously mentioned attributes. [4]

# 4. Methodology

According to Hevner et al. [62], information system (IS) research is operating in the field of action of its environment (s.a. business organizations) and the knowledge base (foundations and methodologies). While the environment provides the relevance with their business needs, the applicable knowledge of the knowledge base defines the rigor. IS research on the one hand allows the application of its results in the respective environment while enriching the knowledge base on the other hand. Within this framework, two (complementary) paradigms influence the research cycle: behavioral-science and design-science. Design-science research is concerned with building of artifacts and their evaluation while behavioral science develops and justifies theories. [62]

Given the case study companies' need for a solution in the customer service channel navigation space without extensive training data available, the chosen approach for this thesis is leaning on design-science research specifically. To understand customer choice determinants, a literature review is conducted and a workshop and interviews at the case study company are held. This yielded decision factors which are then in turn used as input to the developed LLM-based channel suggestion agent and guided the information collected in a user survey for evaluation. Hevner et al. [62] outline seven guidelines for design science in information systems research which they provide to facilitate the understanding of the basis for effective research in this specific area. Despite the fact that attending to these guidelines in some form or another is considered obligatory for complete research, the authors advocate for a liberal use [62]. Table 4.1 outlines the guidelines (*design as an artifact*, *problem relevance*, *design evaluation*, *research contributions*, *research rigor*, *design as a search process* and *communication of research*) along with their original descriptions. To show compliance with the proposed design-science research framework, the following will detail what has been done to address each guideline individually.

*Design as an artifact*. Key goals of the research project are both the compilation of customer service channel choice determinants (factors) and the process and creation of an LLM-enabled decision algorithm for finding appropriate channel suggestions. Hevner et al. [62] argue that also the "methods applied in the development and use of information systems" [62] are considered IT artifacts for this purpose. Hence, the process for obtaining both training data and prompt input also falls within this scope.

*Problem relevance* is indicated by the case study company which aims to harmonize a great customer journey with optimal resource allocation in its customer service. Liu et al. [4] point to the fact that due to a strong user bias towards call hotlines in customer services, both the

| Design Science Research Guidelines | |
|---|---|
| **Guideline** | **Description** |
| Guideline 1: Design as an Artifact | Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation. |
| Guideline 2: Problem Relevance | The objective of design-science research is to develop technology-based solutions to important and relevant business problems. |
| Guideline 3: Design Evaluation | The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods. |
| Guideline 4: Research Contributions | Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies. |
| Guideline 5: Research Rigor | Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact. |
| Guideline 6: Design as a Search Process | The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment. |
| Guideline 7: Communication of Research | Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences. |

Table 4.1.: DSR Guidelines [62]

customers and the company face drawbacks. From the end user perspective this channel means longer waiting times and the company faces non-optimal usage of their channel spectrum. Solutions based on extensive rule sets are both in-flexible and often unaware of the conflicting goals of customer happiness and business resources. [4]

*Design evaluation*. The environment of the case study company currently does not feature a system for automated and personalized channel navigation nor was any data collected for this purpose. Hence, there is no benchmark system available. Consequently, during this thesis, data was collected for this specific task through a user survey facilitated and executed by the case study company. This was then analyzed and transformed into a data set usable for development and testing of the LLM-powered channel decision agent. Performance was then reported using various classical classification metrics (F1-score etc.) and literature inspired custom scoring (see 8.1.1) and compared with learning-based approaches.

The *research contributions* of this thesis are threefold. Firstly, the research on customer channel choice determinants is enriched with a more recent and focused (customer care and service) overview of factors influencing customer channel choice present in the literature. This is extended by the inside-view of the experts from the case study company. Secondly, this research project yields a software artifact for predicting individual channel preferences based on LLM-technology operational with little training data in the realm of customer service.

This includes detailed evaluations regarding the usefulness of detailed customer situation descriptions in the LLM input and benefits of company specific channel descriptions in the prompts. Thirdly, approaches for generating human-level reasons for channel choices are investigated.

*Research rigor* is upheld by basing the factors of customer channel choice on previous research and input from relevant stakeholders. A structured literature review is conducted to identify all relevant articles related to decision factors which is then refined through interviews in the case study company to fit the results to the given situation and processes. Influencing factors on the prompt and input for the LLM-agent were independently tested for their impact on task performance and the reported results for an optimized prompt were based on strictly overlap-free datasets for optimization and testing to avoid over-fit and bias.

*Design as a search process*. The process of finding a suitable solution started with the definition of the problem. This was done in a step-by-step fashion as the most pressing issue of the company was uncovered by iteratively discussing problems and potential solutions in call routing and channel navigation. The eventually stated goal was a solution for predicting customers channel preferences based on to be determined factors. Restricting the selection of tools for this purpose was the desire to work without extensive training data, privacy and compliance considerations and the lack of an existing solution as comparison. The provided solution thus uses a compliant LLM working on input factors that were determined through literature research and a company internal workshop and interviews. The evaluation of the solution is based on data collected through resources available to the company.

*Communication of research*. Each step of the development of the solution, including technical details, is documented in this thesis to ensure reproducibility. Furthermore, managerial implications are derived for a less technology-oriented audience to inform about the necessity to allocate resources for an implementation in production setting.

## 4.1. Case Study

The context in which this research is conducted is provided by a large European insurance company. To reduce the organizational complexity of the problem (channel navigation), the scope is narrowed to one specific department and country. The examples and structures thus pertain to the German branch of damage claims and customer care.

# 5. Relevant Decision Factors

There has been a large amount of research dedicated to examining various aspects of the ever evolving channel landscape that connects consumers and end users with companies [63, 64, 65, 16]. In telephony especially, human call agent touch points are ressource intensive and provide limited scalability with a given set of human operators. Ideally, not each person calling the hotline must be served by this specific channel eventually but could be asked to proceed in a different channel such as chat or online self-service.

In the context of this thesis we specifically aim at combining the customer experience with the company's perspective in customer service channel delivery. To this end, we conduct a workshop and interviews within the case study company to analyze what determines optimal customer channel choice from their point of view. To build on existing research, we additionally undertake a structured literature review to further source customer decision factors. Understanding the decision factors that drive customer channel choice in the first hand is crucial to investigate alternative channel pathways without ruining the customer journey.

In addition to an improved understanding of the overall customer behavior, this yields a framework for developing scenarios that can be used for development and testing of respective solutions in the domain. In terms of building machine learning classifiers and predictors, the resulting list of channel decision determinants also provide a basis for identifying relevant features.

## 5.1. Company Perspective

To include the company specific perspective on the situation, we set up one workshop with five participants and three subsequent one-on-one interviews (see Table 5.1). The goal was to let relevant stakeholders voice their opinion. In close collaboration with the customer care department we were working with, we reached out to experts in the scope of the insurance product we were focusing on. The primary way of contact establishment was through network connections and planning interviews for identifying the appropriate set of participants. All experts that were not able to join the workshop due to scheduling conflicts agreed to a follow-up one-on-one interview to collect their opinion. In both formats, we specifically asked what factors can or should play a role in deciding the optimal channel for a customer. While some of the mentioned factors were business related (such as process cost, ...) the vast majority of answers were focused on the customer and their perspective. An important

| Position | Experience (years) | Participation |
|---|---|---|
| Capacity Manager | 20 | Workshop |
| Capacity Manager | 25 | Workshop |
| Process Lead | 18 | Workshop |
| Project Lead | 5 | Workshop |
| Product Owner | 18 | Workshop |
| Product Owner | 5 | Interview |
| Product Owner | 9 | Interview |
| Business Owner | 20 | Interview |

Table 5.1.: Workshop and interview participants

observation from the interactions was the strong desire of all participants to make sure that the customer journey is as good as it can be and that it should be prioritized over business interests such as cost considerations.

The 90-minute workshop was set-up in a hybrid mode: both virtual (video conference) and on-site participation was possible. The 30-minute interviews were all conducted online via a video conferencing tool. During the sessions, a MIRO[1] board was shared with all participants where the factors that were deemed relevant for channel choice could be collected on note-like objects. Following the raw collection of factors through the brainstorming sessions on the MIRO board, all of them were documented using Microsoft Excel. Duplicates or very similar topics (shared concepts) were grouped together for the workshop and each interview. The aggregated results together with the factors from the literature review can be seen in Table 5.3.

## 5.2. Literature Research

Wolf and Steul-Fischer [17] have created the most extensive and to our knowledge presently most recent systematic literature review in the field of determinants of customer channel choice. They do however not narrow their focus on a specific part of the customer journey or area of product / service. Instead, they aim to combine existing knowlegde about the factors that drive customer channel choice in a mulit- and omnichannel environment across all sectors and customer journey stages. This yielded 66 factors (divided into five main and 14 sub-categories) from 128 papers. [17]

In this part of the thesis we aim to extend their research until the present day (October 2024) and focus on the customer care and service application areas specifically, because this fits the purpose of the case study company. Therefore we conducted a systematic literature review to pinpoint relevant factors for customer channel choice in a multi- and omnichannel en-

[1]https://miro.com; popular collaboration platform

vironment currently present in literature that focuses on customer care and services in general.

### 5.2.1. Methodology

Since we accept the overall premise of the research of Wolf and Steul-Fischer [17] but would like to narrow the results down to a more topic focused (customer care and service) list of decision factors and more recent findings, we follow a similar approach for the structured literature review.

Amongst others, Synder [66] provides a comprehensive guideline for such reviews which is in principle also adapted by Wolf and Steul-Fischer [17]. Accordingly, the process of a structured literature review consists of four phases: designing the review, conducting the review, analysis and writing [66]. An important part of designing the review, is to clarify why, for whom and how this review is conducted [66]. The reason for doing the literature review in the first place is the desire to synthesize and leverage existing research in the field of customer channel choice factors. In terms of the audience ortientation, the review is tailored towards usage within the case study company to complement factors already considered in existing processes. The "how" of the review is determined by the search strategy which starts with the search string. This string inlcudes important terms conncted with Boolean operators. Here, we used the search string from Wolf and Steul-Fischer [17] as basis and refined it for the purposes of this thesis. This meant narrowing down the mulit- and omnichannel choice to customer care and service application areas. The concrete search terms were adapated based on an initial exploratory search and review of the material from [17]. The following was eventually used: ***("Customer Care" OR "Customer Service" OR Hotline OR "Contact Center" OR "Contact Centre" OR "Client service" OR "Client care" OR "Customer support" OR "Post purchase" OR Engagement OR "Service usage" OR Request* OR Complaint*)*** ***AND*** *(choice OR choose* OR select* OR use OR usage OR utili* OR adopt* OR prefer) AND ("omni channel" OR "multi channel" OR "cross channel" OR "dual channel" OR omnichannel OR multichannel OR cross-channel OR dualchannel)*. The parts in bold indicate the difference to the one used by Wolf and Steul-Fischer [17].

Figure 5.1 visualizes how the review was conducted following the definition of the search string. The methodology is almost identical to the one applied by Wolf and Steul-Fischer [17] who based the search process on the PRISMA guidelines [67]. Three databases were used to query relevant papers by means of the above search string on October 14th, 2024: EBSCO Host, Scopus and Web of Science. Inlcusion criteria were defined as English language only, peer-reviewed journals from the domain of management or business categories. The search period was set to start in the year 2000 until the present day since research on digitalized commercial channels and hence multichannel was scarce beforehand [17, 68, 12]. The initial database search yielded 273 papers (EBSCO Host = 61, Scopus = 119, Web of Science = 93) which included 70 duplicates. For the papers which were published before or in 2021, we cross-checked whether they were included in the final review selection of Wolf and Steul-

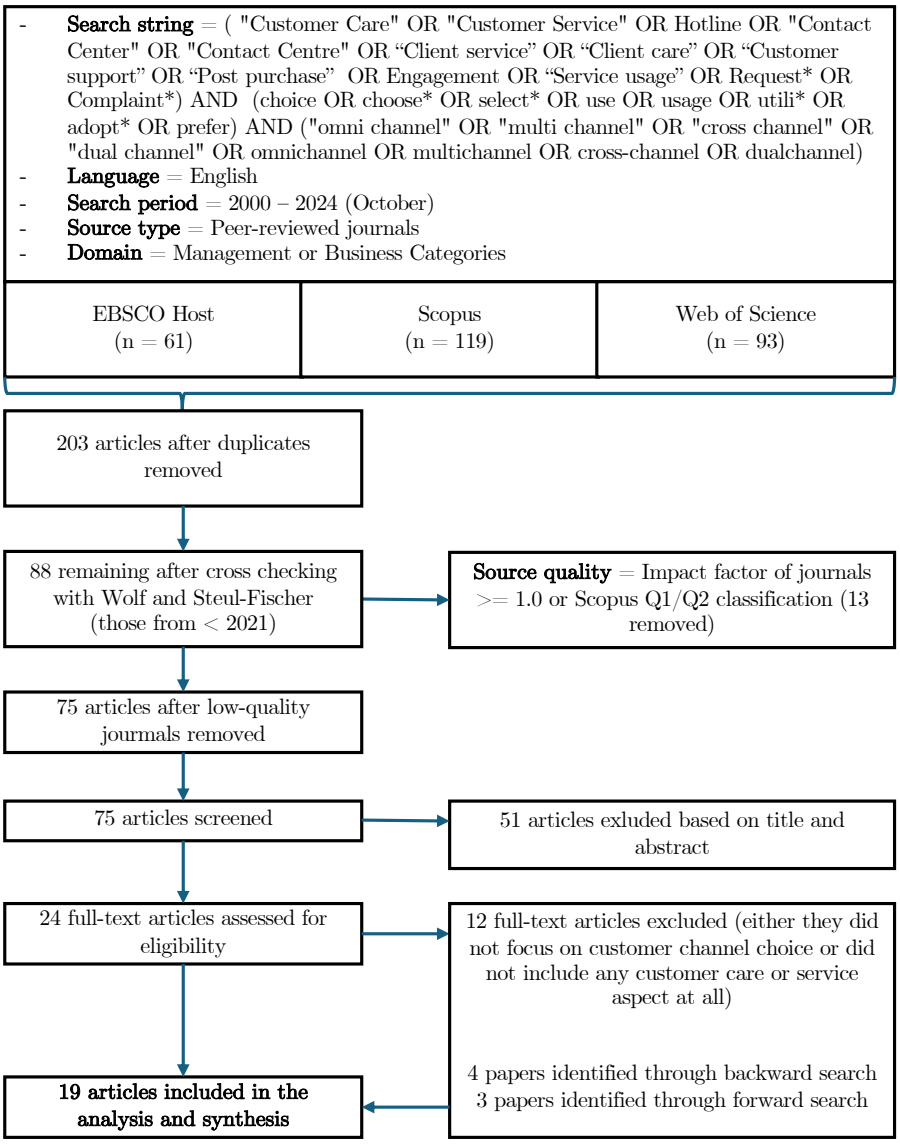



Figure 5.1.: SLR Overview based on [17] and [67]

Fischer [17]. This was possible due to our search string yielding a strict subset of their results because of the additional requirements being added via an AND operator and having the same inclusion criteria otherwise. This left 88 papers for further examination of which 13 were removed due to insufficient journal quality. The quality was assessed by an impact factor greater or equal to 1 (verified via the Journal Citation Report (JCR) by Clarivate [69]) or the declarance of being a Q1/Q2 journal on Scopus if no JCR rating was available [70]. The remaining articles (n=75) were screened by reading the titles and abstracts and evaluated whether they considered choice of (interaction) channel from the customer side specifically and addressed either customer care or service applications. This left 24 papers for a more

detailed analysis of which 12 were futher excluded after reading the entire article. Backward and forward search on this set of articles yielded additional 7 papers.

### 5.2.2. Results

In-depth analysis of the final list of articles (n=19) yielded 47 factors (after merging similar concepts) which were gathered and analyzed using Microsoft Excel. Table 5.3 shows the final list of factors. Each unique factor is shown with a collection of the corresponding papers in which it was analyzed. The shown name is thus a generic term for the different expressions used in the articles. Whenever a term is not clear on its own, a short explanation is added.

Table 5.2: Customer channel determinants identified through SLR

| **Generic Term** | **Explanation** | **Referencing Papers** |
|---|---|---|
| Sex | | [71],[72], [73], [74], [75] |
| Age | | [71], [72], [73], [76], [74], [75], [77] |
| Convenience | Ease and speed for completing tasks incl. waiting time | [71], [78], [16], [76], [79], [74], [77] |
| Channel experience / affinity | Prior experience with the channel technology | [9], [73], [80], [78], [16], [76], [79], [65] |
| Redress seeking | Aiming for some kind of compensation | [71], [16] |
| Frequency of use | | [73] |
| Integration quality | How well are the channels integrated between each other | [81], [9] |
| Type of inquiry | E.g.: service or admin usage | [82], [73] |
| Marital Status | | [72] |
| Information need (degree) | How strong is the need for information? | [82], [77] |
| Venting anger | Does the customer want to release frustration? | [71] |
| Social / personal connection | | [71], [78], [16], [80], [76], [79], [74], [65], [77] |
| Customer engagement level (service usage level) | How many products does the customer have? | [73], [76], [77] |
| Trust in security | | [80], [74] |

Continued on next page

Table 5.2: Customer channel determinants identified through SLR (Continued)

| | | |
|---|---|---|
| Privacy | | [80], [78], [80] |
| Previously used channels | | [83], [16], [77] |
| Mobile Identity | "manifests the importance of mobile technologies in defining and holding individual identity"[81] | [81] |
| Relationship quality | Relationship with company | [83] |
| Individual's assertiveness | Inclination to assert opinions with conviction | [16] |
| Infrastructure | Access to enabling respective hard- and software | [80], [74] |
| Channel awareness | | [80] |
| Openess / willigness to change | | [80] |
| (Hidden) charges | | [80], [76], [84], [74] |
| Region | | [72] |
| State per capita income | | [72] |
| Perceived Necessity | | [80] |
| Perceived Reliability | | [80] |
| Current Satisfaction | | [16] |
| Perceived media richness of channel | How much information can be conveyed? | [85] |
| Culture | | [86] |
| Service complexity | | [76], [74] |
| Product risk | Is it a large (monentary) decision? | [76] |
| Income | | [76], [74] |
| Work-hours | | [76] |
| Ethical stance | Perceived unfairness of upcoming channels due to staff reduction etc. | [76] |
| Company image / Stance | | [76] |
| Range of channel provision | | [76] |
| Personal control | | [84] |

Continued on next page

Table 5.2: Customer channel determinants identified through SLR (Continued)

| | | |
|---|---|---|
| Channel usefulness | | [79] |
| Availability | | [74], [78] |
| Service quality | | [74] |
| Recommendations by others | | [74] |
| Error Risk | | [74], [76] |
| Shopping innovation | | [77], [75] |
| Physical restrictions (PR) | | [65] |
| Claim already filed (insurance) | | [77] |
| Capabilities | Believing in one's own capabilities | [65] |

## 5.3. Synthesis

Following the collection of factors from both industry experts as well as the literature, they were merged and analyzed for overlap. Each factor from each source was considered individually for assessing whether there are similar concepts in the other sources. If so, they were merged while noting that both the company and the literature pointed towards it. Together with the factors that were mentioned exclusively either by the literature or the company, three categories were created to group the factors by source: shared, company- and literature exclusive (see Table 5.3). Incidated in bold are the factors which will be used in the development of the LLM agent to inform about the user and the situation at hand. These particular factors are: *age*, *intent*, *previously used channels*, *infrastructure*, *complexity*, *innovativeness* and *existing claim*. They were chosen for three reasons. Firstly, they are both mentioned by experts in the company and previous studies about customer channel choice behavior, indicating some general agreement about their importance. Secondly, the chosen set is both describing customer characteristics and specifics about the situation, thus differing between customers and claims. Thirdly, the majority is accessible by the company or could be calculated from other input factors.

| Factor Overview | | |
|---|---|---|
| **Only in Literature** | **Mentioned in Both** | **Only in Company** |
| Gender | **Age** | App registration |
| Redress seeking | Comfort | Digital channels used before |
| Usage frequency | Channel affinity | End-customer vs. intermediary |
| Marital status | Integration quality | Patience |
| Information need | **Intent** | Acute danger |
| Engagement level | Venting anger | Customer has all necessary data for process |
| Trust in security | Social / Personal connection | Customer authenticated |
| Privacy | **Previously used channels** | Digital process exists |
| Mobile identity | **Infrastructure** | Customer satisfaction with process |
| Relationship to company | Channel awareness | Process-transparency |
| Personal assertiveness | Openness to change | (Process cost) |
| (Hidden) cost | Perceived necessity (to use other channels) | Prioritization |
| Region | **Complexity** | Customers' uncertainty |
| Income per region | Product risk | Potentially not saying the truth |
| Perceived reliability | Range of channel provision | Time already spent in channel |
| Current satisfaction | Channel usefulness | How visible is the phone number vs. other channel entries |
| Perceived media richness | Availability | Pre-filled data available |
| Culture of customer | Service quality | Customer vs. family/... |
| Income | Error risk | Multiple contact attempts in same channel |
| Working hours | **Innovativeness** | Language barrier |
| Ethical considerations | **Previous claim** | Multiple conctact attempts across channels |
| Company reputation | | (Upselling potential) |
| Personal control | | (Conversion rate) |
| Recommendation | | |
| Physical constraints | | |
| Cognitive abilities | | |

Table 5.3.: Synthesis of factors based on source overlap. Factors in bold will be used for the developed LLM agent. Those in brackets are not relevant for the customer choice.

# 6. Agent Development

In the following, we detail the development and decision process for building the LLM-agent for channel navigation and reason prediction in the context of the case study company. For the evaluation of the solution, please refer to Chapter 8.

## 6.1. Motivation

The first problem at hand – deciding the correct customer prediction for each communication channel – is a typical classification problem. Addressing this with machine learning algorithms usually requires some sort of training data. The goal of the desired solution in this case though is to be independent of available data sources and should also work in data scarce environments. Building a deterministic business logic based on handwritten rules requires full coverage of all possible scenarios and scales badly to other application areas (different business units or companies) in addition to requiring a comprehensive understanding of the decision process on the customer side. Instead, we investigated how well Large Language Models perform in this area of customer care with little data available about the customer preferences a-priori. For the second problem – predicting reasons for the channel choice – we can harness the core capability of LLMs: generating text. By inquiring about the reasons for a certain decision and potentially derivate a reason from the user information alone, far more insight could be generated about a customer request than the preference only.

## 6.2. Channel Preference Prediction

### 6.2.1. Prompting Approach

#### System Prompt

Given the goal of a scalable solution in data-scarce environments and API-only access to the LLM, the remaining development and optimization possibilities are mainly targeted towards prompt engineering. Prompts refer to "textual instructions and examples of [...] desired interaction"[87] which are prepended to the input [87]. Prompt engineering refers to the strategic designing of prompts to direct the output of the model [88]. There is a plethora of different prompt engineering techniques available. Sahoo et al. [88] have collected and categorized 29 distinct variants covering – among others – topics such as hallucination reduction, reasoning or code generation. The categories of highest interest for this work are "New Tasks Without Extensive Training" [88] and "Reasoning and Logic" [88]. The first one covers Zero- and Few-Shot prompting which rely on the internal knowledge of the LLM and

the addition of some exemplary input-output pairs respectively. For the reasoning techniques, we selected Chain-of-Thought (CoT) and Tree-of-Thoughts (ToT) prompting which try to induce a reasoning process. [88]

The initially designed prompt was hand-crafted but loosely based on a prompt from Anthropic's[1] prompt library for a review classifier [89]. It contains details regarding the names of currently available communication channels, the names of the labels that should be assigned to each channel and instructions regarding the desired output format (for details, see Appendix B). Listing 6.1 showcases the expected result format (JSON) of an LLM call with <*Label*> being one of either *preferred*, *acceptable* or *undesired*[2]. Multiple channels with the same label were possible because the eventual test data features this as well.

Listing 6.1: Desired output structure of the LLM agent

```
{
    "Call": <Label>,
    "Chat": <Label>,
    "Self-Service Web": <Label>,
    "Self-Service App": <Label>,
}
```

To aide the process of finding a good prompt, a series of variants were created with all of them containing above mentioned information (albeit in different forms). Four different variables were tested for their influence on the performance on the end-task: *type of instructions/prompts*, *inclusion of few-shots*, *channel description type* and *inclusion of reasons*. The *type of instruction* referred to the wording of the overall task description within the prompt. The baseline type was the one which was hand-crafted based on the prompt library. One alternative was a text generated through a meta-prompt[3] provided by OpenAI. The goal of this prompt is to incorporate best practices for prompt creation in the generation of a new prompt based on an existing one [90]. Hence, the hand-crafted prompt was given as input for the meta prompt which yielded an LLM-generated prompt to serve as an alternative. The other alternatives were CoT- and ToT-prompting. For CoT-prompting we specifically addressed zero-shot-CoT [91] since we did not want to interfere with the few-shot prompting. In zero-shot-CoT no concrete example needs to be provided but only the additional sentence "Let's think step by step" [91] before the answer [91]. As we additionally tried a zero-shot-CoT variant in which we explicitly name a series of steps to analyze the user input, the former (original) zero-shot variant is called *short* zero-shot-CoT here. For Tree-of-Thought-prompting, we used a version from Hulbert [92] which tries to induce multiple reasoning paths and self-evaluation in one promp [92]. The details for the prompts and steps can be found in Appendix B. Furthermore, the *inclusion of few-shots* enriched the given instructions with input-output examples – here:

[1]Anthropic is an AI safety and research company providing the *Claude* LLM family. https://www.anthropic.com
[2]The respective German labels were *Mein präferierter Weg*, *Nicht ideal aber akzeptabel* and *Würde ich nicht nutzen*. The English labels were chosen based on their translation and succinct token representation
[3]https://platform.openai.com/docs/guides/prompt-generation, Last accessed: 25/01/2025

list of factors with user and situation information and expected JSON. The considered options were none (zero-shot), three and five shots. The list of examples for the few-shots were picked from the survey data (see 7) to include each available intent (liability case, accident, ....) at least once. The other factors were varied as much as possible between the examples. For the 3-shot version, the intents were chosen to be *break-in*, *pipe damage* and *accident* as they were deemed the most distinct among the available options. To inform the LLM about the details of each available channel (Call, Chat, Self-Service Web, Self-Service App), experts from the case study company were interviewed for their input (for details, see 6.2.2). The way this information was conveyed in the prompt is referred to here as *channel description type*. Four different variants were tried: bullet-point list (original), plain text, JSON format and no descriptions of the channel at all. Finally, the *inclusion of reasons* collected from the conducted survey (for details, see 7.3) enabled the provision of frequently mentioned reasons for preference choice by the users in the prompt. For each combination of channel and preference label (e.g.: Call & preferred), the generated summary of the most frequent reasons was given.

For each variable, their influence on the performance was tested separately to learn more about helpful prompt construction. The results can be found in Chapter 8.

**Input**

The input (user prompt) consists of a dictionary style string representing various factors about the user and the situation. The concrete selection of factors is based on the literature review and internal interviews with the case study company regarding customer channel choice determinants (for details, see 5.3). In the current version of the agent – in addition to the user utterance itself – seven pieces of information can be provided as input: *age*, *innovativeness*, *available devices*, *previously used channels*, *intent*, *complexity*, *existing claim*. The first four factors refer to the customer while the last three are describing the issue for which the customer is approaching the company. Innovativeness is given on a 5-point Likert scale from very high to very low and the available devices and previously used channels are lists of the respective items. The intent is the category of the claim (e.g.: accident), complexity is either *high* or *low* being an attribute of the insurance situation at hand and existing claim describes whether this communication is about something that has already been mentioned to the company or not.

Similarly to the approach for the system prompt, these factors were individually tested for their influence on the model performance. To this end, more information was incrementally added to the user utterance only to observe potential differences in the respective outcome. The first addition to the user utterance were the three situational factors: intent, complexity and existing claim. This was then considered the baseline as it roughly matches the descriptive information the user saw in the survey scenarios. Following this, the remaining four customer related factors: age, innovativeness, available devices and previously used channels were independently added to the baseline and the respective model performance was observed.

### 6.2.2. Collection of Expert Input

To adapt the prompt to the reality of the case study company, three subject matter experts were interviewed to receive detailed information about the communication channels. In these semi-structured interviews – all conducted online with a video conferencing tool – four questions were discussed: *Which information is necessary to operate in the given channel?*, *In which cases should this channel be used and why?*, *What are the advantages of the channel?* and *What are the disadvantages of the channel?* The results were noted with bullet points and shared with the interviewees afterwards for potential corrections. This information was then used verbatim for the prompts in the format *question: answers in bullets*. To investigate the influence of different textual presentations of the information, two additional variants were created. The JSON format was identical in the wording of the bullets but each question was a separate JSON object (under the keys: "information", "use-cases", "advantages", "disadvantages"). Each object then had a description which was the exact wording of the question and details which was a list containing the bullets. To provide another alternative, a plain text version for each channel was created. This was done by asking OpenAI's GPT-4o model to provide a concise paragraph containing the information from the bullet-points.

### 6.2.3. Technical Realization

Due to company rules and restrictions, the only available LLM setups includes the models from OpenAI through the corporate Azure API. The models are easily accessible by their Python bindings. System and user instructions are passed to the respective Python function. To ensure a high degree of reproducibility (despite inherent probabilistic traits), the seed parameter for the model was set to 42 and it's temperature to 0.00001. Lower values (closer to zero) for the temperature will make the model more deterministic although setting it to zero will lead to an automatic and unknown increase in the background [93]. The value of 0.00001 was chosen because of its existence in the openly available source code for the comparable DeepSeek V3 model[4]. The result is then parsed into a Python object with a custom parsing function. The potentially available structured-output[5] functionality which would allow the result of the API call to be a syntactically correct Python object was not used to decrease the reliance on OpenAI and open up the possibility to switch to other providers without the need for large modifications. The parsing function first extracts the JSON string from the result string. Afterwards, the expected channel-preference pairs were collected and stored in a Python dictionary. To account for potential typos in either the channel or the preference label, Python's difflib library was used to find the most likely target word.

To showcase the ability of the LLM agent and the feasibility of its integration into the existing pipeline, two demo applications were developed (see Figure 6.1). For demonstrating the functionality of the agent, a Streamlit[6] demo application was built. Here, various factors

[4]https://github.com/deepseek-ai/DeepSeek-V3/blob/main/inference/generate.py, Last accessed: 22/02/2025
[5]https://platform.openai.com/docs/guides/structured-outputs, Last accessed: 26/01/2025
[6]Simple Python app framework, https://streamlit.io

about the situation and user (intent, age, previously used channels, ...) can be specified to view the respective result of the LLM agent given this input. This is meant to interactively test the developed application and derive insights for the prompt in the future. To integrate the LLM agent into the call channel as incoming communication mode, Cognigy[7] is used. The software provides (call and chat) flow management and allows to plug in arbitrary services and additional workflows. To make the channel suggestion engine (LLM agent) available here, a dedicated Python web server (for development of the prototype, flask is used) hosts the respective API-calling functionality. The Cognigy call flow includes an automated authorization of the customer and NLP-enabled intent recognition (from a fixed set of intents). Once the intent has been recognized, the flow queries the web server running the channel suggestion engine with the intent and any number of available information about the conversation (such as the customer identifier). Here, further information could be queried from the core database of the company based on the identifier and passed to the LLM. The list of channels for the LLM to choose from is dynamically adapted based on the concrete intent to only include viable options for the concrete use case. During the course of this thesis, a prototype of this setup was built that includes the flask web server with the LLM agent and the integration into an exemplary Cognigy flow to showcase a blueprint for production integration. The connection to the companies' core database with the provision of actual customer data was not realized due to compliance and time reasons. Instead the necessary user information is mocked for demonstration purposes.

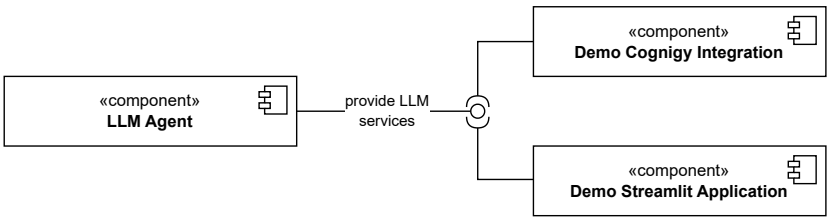



Figure 6.1.: High level overview of the involved components for the demo applications

## 6.3. Reasons Prediction

The availability of free text answers for the questions why the users chose a certain label for a channel opened the opportunity to investigate how well the LLM could generate these. To this end, two approaches were tested. First, the LLM was given the input factors (user and situation information) and the respective decision the user made for this situation. Since each survey participant was only asked for the reasoning behind one channel, the LLM is then asked to generate a justification for this decision from the perspective of the user. In a second approach, the decision of the user was additionally masked (effectively only providing the

[7] https://www.cognigy.com

factors) to prompt the LLM to first provide a channel prediction and then answer the question why this was suitable here.

For the first approach (explain user decision), the LLM was given the following prompt: *You are an analyst in an insurance company. Your job is to predict why a customer labeled a certain communication channel with a preference. The possible channels are: Call, Chat, Self-Service Web, Self-Service App. The possible labels are preferred, acceptable, undesired. You will be given some information about the user and his or her problem along with the respective decision for a certain channel. Respond with a short explanation why this was chosen. Answer in German. Speak as if you were the customer who asked this question. Be colloquial. Restrict your answer to one very short sentence maximum.* The last four sentences were necessary to guide the results towards the style of the answers provided by the users. The input was then given in the form: *User info: <user_info>. The user chose the label <preference> for the channel <channel>* with user_info being the factors and preference and channel the respective preference label and channel name that was chosen here.

In the second approach the LLM was first queried for the classification of the channels based on the user input with the regular prompt. Following this, a second query was created to retrieve the reason. This prompt started with the exact same first four sentences of the prompt from the first approach. To adapt to the new setting, it continues as follows: *You will be given the information about the user and his or her problem along with the respective decision you made for him or her. Respond with a short explanation why you chose the label <preference> for the channel <channel>. Answer in German. Speak as if you were the customer who asked this question. Be colloquial. Restrict your answer to one very short sentence maximum.* The respective input then was: *User info: <user_info>. Your decision was: Call (<label>), Chat (<label>), Self-Service Web (<label>), Self-Service App (<label>).*

# 7. Data Collection

To develop and test the envisioned solution, test data is required. However, since no such data currently exists, test cases need to be generated. Here, tests pertain to a certain situation in which a customer calls (or contacts) the company and the respective preferences regarding channels in the answering of the issue – with the preferences serving as the labels. As no such data can be collected from the company's live systems for compliance reasons, the situation shall be presenteted to end users in form of a survey with corresponing questions for retrieving the preferences. This process leans on experimental vignette methodology, which involves providing participants with carefully synthesized, realistic scenarios for evaluation [94].

## 7.1. Scenarios

To structure and guide the process of creating these scenarios, previously identified factors (see 5.3) are used to differentiate between cases and ensure that as much relevant information as possible is considered. Instantiating a scenario involves selecting values for these factors. Realizing all possible combinations of factor values is infeasible due to the magnitude of labeling required. To address this, a selection of relevant and sufficiently broad situations were chosen based on their importance for the case study company. Hence the scenarios were related to the liability, accident, pipe-damage, break-in and lost-key insurances. For each of the five topics, four variations were hand-crafted based on the level of complexity (low, high) and whether this has been a new claim or already reported to the insurance. All of these were exemplary situations not based on existing customer data. Each scenario starts with the request to imagine oneself in the situation in which one is calling the insurance company with the mentioned issue but has not been connected to an employee yet. Table 7.1 shows the four scenario variants based on the lost-key insurance. The difference in complexity in this case is realized by varying the amount of damage and the associated (perceived) level of administrative effort. The full list of scenarios (20 in total) can be found in Appendix A.

## 7.2. Survey

To translate scenarios into test cases, they must be tied to user-based channel preferences. To this end, the case study company offered logistic support in conducting a survey based on the identified scenarios and user-related factors to generate insights into the channel (switching) preferences of customers. N=709 people, identified and contacted by the case study company,

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen komplizierten Schadenfall gemeldet. Sie haben einen Schlüssel der zentralen Schließanlage Ihres Büros verloren. Der Verlust zieht sehr hohe Kosten mit sich. Jetzt wollen Sie sich bei Ihrem Versicherungsunternehmen über den aktuellen Bearbeitungsstand erkundigen. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen simplen Schaden gemeldet. Sie haben letzte Woche den Wohnungstürschlüssel in Ihrer Wohnung vergessen und haben sich ausgesperrt. Deshalb mussten Sie einen Schlüsseldienst beauftragen. Jetzt wollen Sie sich bei Ihrem Versicherungsunternehmen über den aktuellen Bearbeitungsstand erkundigen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadensabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben einen Schlüssel der zentralen Schließanlage in ihrem Büro verloren. Der Verlust zieht sehr hohe Kosten mit sich. Diesen komplizierten Schaden wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben letzte Woche den Wohnungstürschlüssel in Ihrer Wohnung vergessen und haben sich ausgesperrt. Deshalb mussten Sie einen Schlüsseldienst beauftragen. Diesen Schaden wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table 7.1.: Exemplary scenarios for the intent: lost keys

received a questionnaire (based on our questions) via e-mail and were presented with three randomly selected scenarios. The survey was originally conducted in German. Questions and answers are reported in English here for better readability. Each scenario had six associated follow-up questions to retrieve information about contact/communication channel preferences:

1. You want to notify the insurance company about this situation. This can happen via a call but also via different communication channels. There is also the possibility to switch communication channels during one claim. How do you assess the following communication channels for reporting the previous situation? The listed channels are: call/call-back, chat, website, app. Each channel can be given one of three categories: "preferred", "not ideal but acceptable", "would not use".
2. Please justify why you chose <preference> for channel <channel>?
3. For the description of the situation you were given various information. Please rank the information according to their importance (complexity, intent, previous claim).
4. How would you describe the situation if asked: "What can we help you with?" (relating to the aforementioned scenario)
5. Was there any information missing to make a good decision about the communication channel you are choosing?
6. Which information was missing?

Question 1 allowed for several channels to have the same labels. For question 2, exactly one of the four answers was chosen to not tire-out the participants. The decision was made by sampling with a 40% probability an answer that was either "preferred" or "would not use" to obtain more pointed answers.

The resulting answers were collected in a dataset which comprises of 2127 individual answers. In addition to the questions regarding the situation, a series of self-assessments for example regarding one's innovativeness or waiting time patience were asked to build a customer profile and provide values to the aforementioned factors. All available variables in the resulting data set along with their type and possible values are recorded in Table 7.2. A rough characterization of the involved participants as presented in Figure 7.1 and 7.2 shows that almost all age categories between 23 and 87 are present with higher density distributed around 40 and 60 (mean: 53.3, median: 54). When it comes to self-assessing the personal innovativeness, the participants seemed to be leaning towards being open to use new technology and consider it easy to use new software. The interest in artificial intelligence (AI) is less distinct with more people resonating with being absolutely not interested in AI than the opposite.
Each scenario (intent + complexity + existing previous claim) was answered between 104 and 109 times[1]. The answer distribution for each channel (see Figure 7.3) reveals that only a small

[1]The counts are not exactly equal due to random sampling of the situations in the creation of the questionnaire

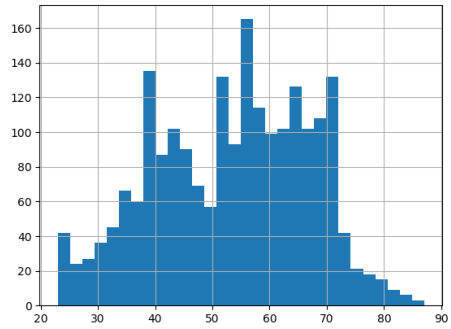
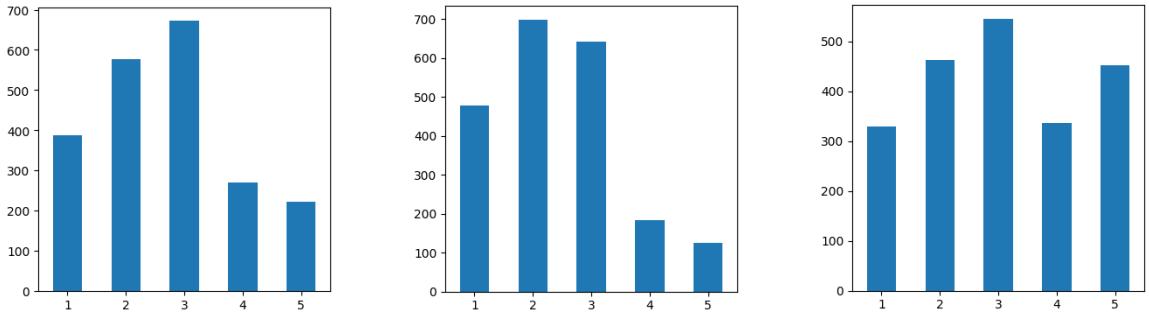



Figure 7.1.: Age distribution




(a) Proneness to use new technology

(b) Ease to use new software

(c) Interest in AI

Figure 7.2.: Participant answer distribution for a 5-Point Likert Scale. 1 means full agreement, 5 means full disagreement, 3 is neutral

fraction of users would not consider using the call channel at all with around 70% choosing it as a preferred channel. This is in stark contrast to the answers given for the chat and app channel. Although the app channel is the only one, where more people would not use it than any other option, both for the app and chat, less than half of the number of people who would not use it would consider it a preferred option.

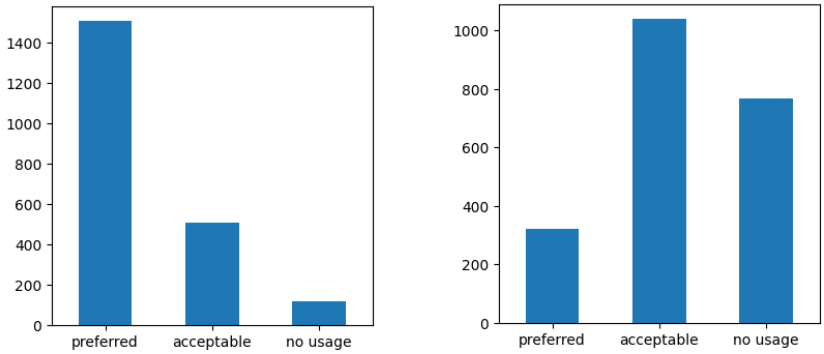
## 7.3. Preference reasons

The opportunity for survey participants to state their reasons for choosing a particular preference label for a channel opens up the possibility of tapping into users' detailed motivations regarding their communication channel considerations in the context of insurance claims. For each channel and preference, there are a varying number of free-text answers on why the respective label was chosen for the channel. This can be traced back to the different counts of preference labels per channel and im-balanced sampling of questions which was applied to bias the number of answers towards the labels *preferred* and *would not use*. To find patterns and

| **Evaluation data set** | | |
|---|---|---|
| **Variable** | **Type** | **Variable Level** |
| Age | Numeric | |
| Previously used channels | Categorical | personal visit at broker, personal visit at home, video-call with broker, call with broker, hotline, e-mail, WhatsApp, chat with human, chatbot, customer portal, app, sms, web-form |
| Maximum waiting time hotline satisfied | Numeric [min] | |
| Minimum waiting time hotline dissatisfied | Numeric [min] | |
| Proneness to use new technology | 5-Point Likert Scale | 1-5 |
| Interest for AI | 5-Point Likert Scale | 1-5 |
| Ease to use new software | 5-Point Likert Scale | 1-5 |
| Registration at customer portal | Binary | True, False |
| Customer type | Categorical | Proprietary customer types (4) |
| Available devices | Categorical list | tv, pc, laptop, tablet, e-book, phablet, smartphone, regular mobile, gaming console, voice assistant, media receiver, streaming box, smartwatch, hybrid pc |
| Intent (type of damage) | Categorical | (liability, lost key, accident, break-in, pipe-damage) case |
| Complexity of claim | Categorical | Low, High |
| Existing previous claim | Binary | True, False |
| Preference for call channel | Categorical | preferred, not ideal but acceptable, would not use |
| Preference for chat channel | Categorical | preferred, not ideal but acceptable, would not use |
| Preference for web channel | Categorical | preferred, not ideal but acceptable, would not use |
| Preference for app channel | Categorical | preferred, not ideal but acceptable, would not use |
| Reason for choosing preference of channel X | Free text | |
| Ranking of relevance of given information | Ranking of options | complexity, intent, previous claim |
| Description of situation in own words | Free text | |
| Was there necessary information missing to decide the preference? | Binary | |
| Information that was missing | Free text | |
| Not used: state, job type, school education, household size and number of children | | |

Table 7.2.: Evaluation data set

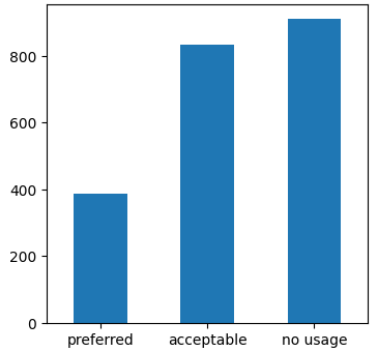
re-occuring topics in the reasons (specifically in the training data subset, see Section 8.1.1), the texts were clustered using hierarchical clustering[2]. Hierarchical clustering was chosen over other clustering mechanisms like K-Means to avoid having to pre-define the number of resulting groupings. Exploratory experiments and investigation of the dendrogram suggested 0.54 to be a good fit for the threshold of the distance criterion of the applied algorithm. The pairwise comparison of text values was made possible by embedding each reason into an array of floating point numbers using the AzureOpenAI embedding model *text-embedding-3-large* and subsequently applying the cosine function between two arrays to determine the similarity.



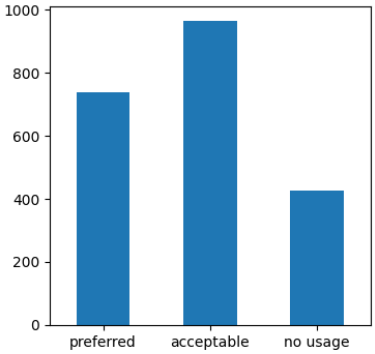

(a) Call




(b) Chat




(c) Self-Service App




(d) Self-Service Web

Figure 7.3.: Participant answer distribution for each channel when asked for their communication preferences in various situations. Labels shortened for visual reasons.

The resulting clusters for each channel-preference combination were then sorted by the number of reasons they contained, assuming the size of a cluster indicated its relevance for the overall understanding of important topics. To aggregate the theme of each prioritized cluster, the GPT-4o LLM was used to generate one descriptive sentence based on all the different reasons in the cluster. The final prompt used for this task was: *You are a business*

[2]Implemented using the Python package SciPy and the methods `cluster.hierarchy.linkage(X, ''average'', metric=''cosine'')` and `cluster.hierarchy.fcluster(Z, threshold, criterion=''distance'')`

| Channel & preference | Random example | Generated summary |
|---|---|---|
| Web & preferred | "It's the fastest" | "Speed, Simplicity, No waiting times, Convenience, 24/7 availability, uncomplicated nature" |

Table 7.3.: Exemplary reasons

*analysis assistant for an insurance company. You will be given a list of customer expressions in the context of a customer service center. The expressions are answers for the question why the customers considered the <channel> channel as <preference>. What are the themes that are shared between the expressions? Only respond with the list of the generic names of the theme(s). Be very brief and concise and do not add explanations. The result should look like this: [Theme 1], [Theme 2], [Theme 3], ....* The <*channel*> and <*preference*> tags were dynamically replaced with their actual channel and preference labels to set the correct context. If the output did not match the desired structure, the keypoints were manually extracted (and translated to English if necessary). The answer for the web channel and label acceptable was rejected by the models content policy and thus manually summarized. Table 7.3 shows the resulting description for the *web* channel and the *preferred* label along a randomly sampled reason[3] that was given in this category.

In addition to the free text answers why a channel was given a specific label, survey participants were also asked which of the three pieces of information they considered how important for their decision: *complexity of the matter*, *existence of previous claim* and *intent*. The *existence of previous claim* refers to whether the matter at hand had been communicated with the company already or whether this contact is the first time the insurance company hears about the claim. The result was a ranking of these factors for each situation. The items most frequently chosen for the first rank were:

1. *intent* (41%)
2. *existence of previous claim* (37%)
3. *complexity of the matter* (22%)

This gives rise to the conclusion that from the situational characteristics, the *intent* and the *existence of previous claim* are the key factors for making the channel choice (or preference) decision.

## 7.4. Statistical Factor Insights

The availability of data also motivated the statistical analysis of the correlation between factors and channel preferences. For the following considerations, we focused on the situational factors (intent, complexity and previous claim) as these were the only ones predetermined

[3]Translated from German and slightly modified for privacy reasons

by us in the survey with very similar sample sizes. To achieve this, twelve binary variables were created by combining each channel with each label (call_preferred, call_acceptable, ...) which are set to true if the respective combination holds true for the answer of the user. Subsequently all situational factors were tested for correlation with all dependent variables (call_preferred, ...) using the chi-squared test[4] on the basis of the contingency table of the relation[5]. Additionally, the association strength between the two variables is reported using Cramer's V[6]. According to Akoglu [95], a corresponding value above 0.05 means a weak association while a value above 0.1 indicates a moderate association. Values above 0.15 even hint to a strong association [95]. Since we look at all the calculated results to determine potentially significant associations (by looking through the p-values), we need to correct for multiple testing in order to account for the high number of hypothesis we are testing and reduce the probability of rejecting some true null hypothesis (no association between the variables) by chance [96]. One way of doing so is with the Benjamini-Hochberg method [97]. Thus, we restrict the respective false discovery rate[7] to 0.05 by looking at the adjusted p-values.

Table 7.4, 7.5 and 7.6 show the situational variables that exhibit a significant correlation to the channel being labeled "preferred", "acceptable" or "would not use" respectively. The variable intent has been stratified into binary variables for each intent for more granularity. Additionally, Table 7.7 provides the relevant contingency tables offering insights into the nature of the relationship between the variables. Table 7.4 thus indicates that the intent lost-key and the variable previous damage/claim have a statistically significant correlation with whether the web channel is labeled as "preferred" or not. The share of people preferring the web channel is lower among people who reported a stolen key or have a previous claim (damage) than those who did not. Complexity even seems to correlate with the preferred label for three channels: call, chat and web. People reacting to low complexity situations were more often preferring the chat and web channel than those answering to high complexity situations. Similarly, high complexity situations led to more people preferring the call channel than those with low complexity.

For the presence of the label "acceptable" (see Table 7.5), only the variable complexity and the call channel seem to be significantly associated. Respondents working with low complexity scenarios were more likely to favor the acceptable label for the call channel than those with high complexity scenarios.

The labelling of the web channel with "would not use" (see Table 7.6) seems to correlate with an intent – similarly as does the labelling the channel with preferred. Here it is the intent

[4]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html, Last accessed: 22/02/2025

[5]https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html, Last accessed: 22/02/2025

[6]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.contingency.association.html, Last accessed: 22/02/2025

[7]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.false_discovery_control.html, Last accessed: 23/02/2025

| Variable | Target | p-value | adj. p-value | Cramer's V |
|---|---|---|---|---|
| intent_lost-key | self-service web_preferred | 0,000531516 | 0,011161831 | 0,076350218 |
| complexity | call_preferred | 0,001233206 | 0,014798475 | 0,071092721 |
| complexity | chat_preferred | 0,001899444 | 0,018912399 | 0,068648234 |
| complexity | self-service web_preferred | 1,30711E-06 | 5,48985E-05 | 0,105904148 |
| previous_damage | self-service web_preferred | 0,002225522 | 0,018912399 | 0,067301954 |

Table 7.4.: Associations with channels deemed "preferred"

| Variable | Target | p-value | adj. p-value | Cramer's V |
|---|---|---|---|---|
| complexity | call_acceptable | 0,00049704 | 0,011161831 | 0,07661136 |

Table 7.5.: Associations with channels deemed "acceptable"

accident as opposed to lost-key for the preferred case. There appears to be a higher prevalance of the label "would not use" for the web channel among those dealing with accidents than for those who are not. The existence of a previous damage or claim also positively correlates with the label "would not use" for the web channel. Complexity seems associated accordingly with chat, web and app – for all of which the share of people reacting to high complexity scenarios and labeling the respective channel as "would not use" is higher than for low complexity scenarios.

| Variable | Target | p-value | adj. p-value | Cramer's V |
|---|---|---|---|---|
| intent_accident | self-service web_no_usage | 0,001062861 | 0,014798475 | 0,072442888 |
| complexity | chat_no_usage | 0,000709252 | 0,011915432 | 0,074397029 |
| complexity | self-service web_no_usage | 2,49564E-07 | 2,09634E-05 | 0,113015479 |
| complexity | self-service app_no_usage | 0,002251476 | 0,018912399 | 0,067188828 |
| previous_damage | self-service web_no_usage | 0,0039493 | 0,030158294 | 0,063668684 |

Table 7.6.: Associations with channels deemed "would not use"

| | | self-service web (preferred) | |
|---|---|---|---|
| | | *False* | *True* |
| **intent lost-key** | *False* | 1080 | 620 |
| | *True* | 310 | 117 |

| | | self-service web (undesired) | |
|---|---|---|---|
| | | *False* | *True* |
| **intent accident** | *False* | 1385 | 316 |
| | *True* | 316 | 110 |

| | | call (preferred) | | chat (preferred) | | self-service web (preferred) | |
|---|---|---|---|---|---|---|---|
| | | *False* | *True* | *False* | *True* | *False* | *True* |
| **complexity** | *high* | 275 | 783 | 924 | 134 | 745 | 313 |
| | *low* | 347 | 722 | 881 | 188 | 645 | 424 |

| | | chat (would not use) | | self-service web (would not use) | | self-service app (would not use) | |
|---|---|---|---|---|---|---|---|
| | | *False* | *True* | *False* | *True* | *False* | *True* |
| **complexity** | *high* | 639 | 419 | 798 | 260 | 570 | 488 |
| | *low* | 722 | 347 | 903 | 166 | 647 | 422 |

| | | self-service web (preferred) | | self-service web (would not use) | |
|---|---|---|---|---|---|
| | | *False* | *True* | *False* | *True* |
| **previous damage** | *False* | 658 | 401 | 874 | 185 |
| | *True* | 732 | 336 | 827 | 241 |

Table 7.7.: Contingency tables for statistically significant associations

# 8. Evaluation

The existence of a comprehensive data set opens up the possibility of evaluating the LLM agent on actual customer data. First, the channel preference prediction capabilities will be investigated. This is followed by an examination of the results from the reason prediction.

## 8.1. Evaluation of Channel Preference Predictions

Although the channel preference prediction approach was build with data scarcity in mind, analyzing the performance empirically is crucial for making statements about the usability and potential scalability of the chosen methods in other application areas. In the following, we will report the methodology of the evaluation and subsequently the results.

### 8.1.1. Methodology

#### Data

An important consideration of the evaluation was the treatment of the available data. The data set features a field which contains the users free-text answer to how they would have phrased this scenario towards the insurance. Some survey participants did not properly answer this question in the right format which would have rendered its use in the LLM input useless. Therefore, the data set was manually filtered to exclude these cases from further experiments. This affected 571 (=26.8%) of the data points. The remaining 1556 data points were treated as follows: five examples were set aside for few-shots and the rest was randomly split into roughly 80% (training/tuning) and 20% (test)[1]. Accounting for the fact that each person was responsible for three data points, the split was done on the person instead of the data rows. Resulting from this was a training data set with 1243 data points and a test set with 308 data points. This ensured that the data used in few-shot prompting or tuning was not wrongly biasing the test results.

#### Automation

To automate and simplify the evaluation process, a series of Python scripts were developed. Test cases (scenarios) are provided as CSV files and contain all the necessary information about the situation such as the user factors and the corresponding decision that was made by the user for the channels. An evaluation run then loops through all relevant test cases and the

[1]The split was implemented using the Python package scikit-learn and the method `model_selection.train_test_split(person_ids, test_size=0.2)`

results are collected and aggregated. Each test case execution further produces a log file with all relevant details (model, results, prompts, ...). To speed up the processing time, several different evaluations with different prompt or input strategies could be run in parallel.

**Metrics**

The output structure of the LLM agent gives rise to various possibilities of reporting the scores. One holistic approach was inspired from Rao et al. [59]. It introduces a scoring mechanism that assumes a certain compatibility between answers rather than simply looking for exact matches [59]. This is applicable here since the difference between labeling a channel as *would not usen* and *preferred* is larger than the difference between *preferred* and *not ideal but acceptable*. Consequently, we opted for the following custom scoring mechanisms of expected and (LLM-)suggested labels:

- Exact match: 1
- *Preferred* & *not ideal but acceptable*: 0.5
- Others: 0

The custom scores reported in the following are always the percentage of the actually obtained points compared to all obtainable points. Other applicable performance measures are confusion matrix-based metrics such as accuracy, precision, recall and F1-score. These are be applied on a per-channel basis to compare the correctness of the preference label. While accuracy for example presents itself as an intuitive metric, it shows weaknesses in the case of unbalanced data sets [98]. The Matthews Correlation Coefficient (MCC) is less susceptible to this [98, 99]. The scikit-learn Python library offers implementations for each of the desired metrics[2] with inherent support for the multiclass case (due to three labels per channel). In the case of precision, F1-score and recall, the respective *average* parameter of the scoring function was set to "*weighted*" to factor in to some degree the balance of the given labels.

### 8.1.2. Results

The respective scores for the different input and system instruction variants are reported on the training dataset. A promising candidate from these runs is then selected to be run exactly once on the test set. This allows for an unbiased reporting of the model performance on the test set. Initial test runs with various prompt combinations on the training data set quickly confirmed no relevant difference between the performance of the LLM agent when using either the GPT-4o or GPT-4o-mini model. Hence, the following experiments were run on GPT-4o-mini due to considerably lower cost. For full disclosure, the scores for each channel are reported separately for most evaluations.

[2]https://scikit-learn.org/stable/api/sklearn.metrics.html, Last accessed: 17.02.2025

### User Input

Table 8.1 details the performance metrics of the different input variable combinations across all channels on the training data set. Since two specific situations were flagged and subsequently blocked by OpenAI's content filter due to alleged violence and jailbreak attempt, these were removed from all seven runs. *Only utterance* describes the input which features only the free text answer of the user to the question of how this situation would be described by them. The **Baseline** extends this by the situational factors of intent, complexity and previous claim. Looking at the metrics across all channels, we can observe that the baseline outperforms the simple user utterance as input only in precision and Matthews Correlation Coefficient (MCC). In all other metrics, the utterance variant seems to perform better. Important to notice is here that the MCC score of the utterance variant is very close to zero strongly indicating randomness. For the independently added factors age, previously used channels, available devices and innovativeness, previously used channels seems to contribute slighlty the most to performance on its own although the results are very close to each other with the exception of innovativeness which is scoring worse in all metrics but precision and notably worse for the MCC. The most important result from this experiment is that the combination of **all** factors toghether yielded the best performance across all metrics.

### System Instructions (Prompt)

For the different prompt strategies the addition of few-shots, commonly named reasons and channel descriptions was tested as well as some alternative prompt types. The baseline here is define as the hand-crafted prompt without any additions (for details, see 6.2.1). Table 8.2 shows how the addition of three- and five-shots and inclusion of reasons on top of the baseline prompt performed in comparison. 5-shot prompting produced a better score for each metric when considering the across channel values but interestingly even outperforms the baseline for each metric in the channel specific scores of chat, web and the app. The inclusion of reasons on the other hand showed no significant added value compared to the baseline with only being very slightly above the baseline - but in the same region - for the MCC value (in the across channel metrics). Table 8.3 details the comparison between the baseline prompt and the addition of channel descriptions as provided by the case study company experts. Interestingly, the provisioning of this information (in no form) helps to increase the model performance. The results look differently when changing the prompt type from the hand-crafted prompt in the baseline to alternatives. As observable in Table 8.4, all alternatives except the Tree-of-Thought prompt exhibited better scores for accuracy, recall, F1-score, MCC and custom score than the baseline (averaged across the channels).

### Promising Combinations

Following the independent tests of input and prompt strategies, we combine strong candidates of the prior – namely few-shots and Chain-of-Thought (CoT) prompting – to find a candidate for the run on the test set. As can be seen in Table 8.5, the 1-shot CoT variant perfomed well compared to its 3- and 5-shot alternatives. Especially for the F1-score and MCC it shows

| | | | | Separately Added to Baseline | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Only Utterance* | **Baseline** | *+age* | *+channels* | *+devices* | *+innov.* | **All** |
| **Call** | *Accuracy* | **0,645** | 0,525 | 0,570 | 0,566 | 0,496 | 0,500 | 0,597 |
| | *Precision* | 0,534 | 0,595 | 0,589 | 0,601 | 0,596 | 0,591 | **0,603** |
| | *Recall* | **0,645** | 0,525 | 0,570 | 0,566 | 0,496 | 0,500 | 0,597 |
| | *F1-Score* | 0,570 | 0,545 | 0,575 | 0,578 | 0,534 | 0,534 | **0,599** |
| | *MCC* | -0,027 | 0,095 | 0,096 | 0,104 | 0,081 | 0,069 | **0,116** |
| **Chat** | *Accuracy* | 0,434 | 0,366 | 0,404 | **0,469** | 0,398 | 0,395 | **0,469** |
| | *Precision* | 0,255 | 0,265 | 0,262 | 0,436 | 0,267 | 0,442 | **0,459** |
| | *Recall* | 0,434 | 0,366 | 0,404 | **0,469** | 0,398 | 0,395 | **0,469** |
| | *F1-Score* | 0,319 | 0,301 | 0,318 | 0,377 | 0,318 | 0,315 | **0,382** |
| | *MCC* | -0,021 | 0,030 | 0,011 | 0,047 | 0,034 | 0,003 | **0,054** |
| **Web** | *Accuracy* | 0,276 | 0,332 | 0,327 | 0,355 | 0,335 | 0,335 | **0,397** |
| | *Precision* | 0,363 | 0,403 | 0,397 | 0,427 | **0,454** | 0,409 | 0,444 |
| | *Recall* | 0,276 | 0,332 | 0,327 | 0,355 | 0,335 | 0,335 | **0,397** |
| | *F1-Score* | 0,254 | 0,267 | 0,287 | 0,336 | 0,265 | 0,270 | **0,403** |
| | *MCC* | -0,008 | 0,100 | 0,088 | 0,100 | 0,107 | 0,103 | **0,119** |
| **App** | *Accuracy* | 0,408 | 0,340 | 0,411 | 0,350 | 0,345 | 0,335 | **0,451** |
| | *Precision* | 0,381 | 0,380 | 0,420 | 0,363 | 0,449 | 0,336 | **0,461** |
| | *Recall* | 0,408 | 0,340 | 0,411 | 0,350 | 0,345 | 0,335 | **0,451** |
| | *F1-Score* | 0,376 | 0,290 | 0,359 | 0,324 | 0,293 | 0,284 | **0,427** |
| | *MCC* | 0,011 | 0,054 | 0,107 | 0,039 | 0,069 | 0,043 | **0,164** |
| **Across Channel** | *Accuracy* | 0,441 | 0,391 | 0,428 | 0,435 | 0,394 | 0,391 | **0,479** |
| | *Precision* | 0,383 | 0,411 | 0,417 | 0,457 | 0,442 | 0,444 | **0,492** |
| | *Recall* | 0,441 | 0,391 | 0,428 | 0,435 | 0,394 | 0,391 | **0,479** |
| | *F1-Score* | 0,380 | 0,351 | 0,385 | 0,403 | 0,353 | 0,351 | **0,453** |
| | *MCC* | -0,011 | 0,070 | 0,075 | 0,073 | 0,073 | 0,055 | **0,113** |
| | *Custom Score* | 0,530 | 0,523 | 0,544 | 0,551 | 0,516 | 0,516 | **0,588** |

Table 8.1.: Evaluation scores for the different input variables. Two test cases were removed from the training set for these runs due to OpenAI's content filter policy which flagged one for violence and one for jailbreak. Entries in bold indicate a row maximum

| | | | Few-Shots | | Reasons |
|---|---|---|---|---|---|
| | | **Baseline** | *3-shot* | *5-shot* | *Included* |
| **Call** | *Accuracy* | **0,597** | 0,445 | 0,533 | 0,491 |
| | *Precision* | 0,603 | 0,621 | **0,626** | 0,589 |
| | *Recall* | **0,597** | 0,445 | 0,533 | 0,491 |
| | *F1-Score* | **0,599** | 0,442 | 0,543 | 0,512 |
| | *MCC* | 0,116 | 0,128 | **0,165** | 0,068 |
| **Chat** | *Accuracy* | 0,469 | **0,527** | 0,497 | 0,499 |
| | *Precision* | 0,459 | **0,506** | 0,492 | 0,474 |
| | *Recall* | 0,469 | **0,527** | 0,497 | 0,499 |
| | *F1-Score* | 0,382 | **0,477** | 0,430 | 0,468 |
| | *MCC* | 0,054 | **0,151** | 0,101 | 0,118 |
| **Web** | *Accuracy* | 0,397 | 0,455 | **0,476** | 0,376 |
| | *Precision* | 0,444 | **0,480** | 0,475 | 0,449 |
| | *Recall* | 0,397 | 0,455 | **0,476** | 0,376 |
| | *F1-Score* | 0,403 | 0,416 | **0,442** | 0,356 |
| | *MCC* | 0,119 | 0,137 | **0,151** | 0,129 |
| **App** | *Accuracy* | 0,451 | 0,462 | **0,529** | 0,445 |
| | *Precision* | 0,461 | 0,516 | **0,530** | 0,449 |
| | *Recall* | 0,451 | 0,462 | **0,529** | 0,445 |
| | *F1-Score* | 0,427 | 0,471 | **0,525** | 0,422 |
| | *MCC* | 0,164 | 0,205 | **0,257** | 0,146 |
| **Across Channel** | *Accuracy* | 0,479 | 0,472 | **0,509** | 0,453 |
| | *Precision* | 0,492 | **0,531** | 0,531 | 0,490 |
| | *Recall* | 0,479 | 0,472 | **0,509** | 0,453 |
| | *F1-Score* | 0,453 | 0,451 | **0,485** | 0,439 |
| | *MCC* | 0,113 | 0,155 | **0,169** | 0,115 |
| | *Custom Score* | 0,588 | 0,622 | **0,639** | 0,569 |

Table 8.2.: Evaluation scores for the inclusion of few-shots and reasons. Entries in bold indicate a row maximum

| | | | Channel Descriptions | | |
|---|---|---|---|---|---|
| | | **Baseline** | *Bullets* | *JSON* | *Plain* |
| **Call** | *Accuracy* | **0,597** | 0,463 | 0,478 | 0,448 |
| | *Precision* | **0,603** | 0,593 | 0,590 | 0,586 |
| | *Recall* | **0,597** | 0,463 | 0,478 | 0,448 |
| | *F1-Score* | **0,599** | 0,513 | 0,520 | 0,500 |
| | *MCC* | **0,116** | 0,062 | 0,063 | 0,048 |
| **Chat** | *Accuracy* | **0,469** | 0,451 | 0,454 | 0,437 |
| | *Precision* | **0,459** | 0,454 | 0,418 | 0,437 |
| | *Recall* | **0,469** | 0,451 | 0,454 | 0,437 |
| | *F1-Score* | **0,382** | 0,347 | 0,339 | 0,362 |
| | *MCC* | **0,054** | 0,017 | 0,002 | 0,011 |
| **Web** | *Accuracy* | **0,397** | 0,335 | 0,344 | 0,337 |
| | *Precision* | **0,444** | 0,379 | 0,431 | 0,387 |
| | *Recall* | **0,397** | 0,335 | 0,344 | 0,337 |
| | *F1-Score* | **0,403** | 0,267 | 0,279 | 0,279 |
| | *MCC* | **0,119** | 0,104 | 0,117 | 0,101 |
| **App** | *Accuracy* | **0,451** | 0,360 | 0,358 | 0,356 |
| | *Precision* | **0,461** | 0,420 | 0,382 | 0,396 |
| | *Recall* | **0,451** | 0,360 | 0,358 | 0,356 |
| | *F1-Score* | **0,427** | 0,313 | 0,308 | 0,316 |
| | *MCC* | **0,164** | 0,085 | 0,075 | 0,070 |
| **Across Channel** | *Accuracy* | **0,479** | 0,403 | 0,408 | 0,395 |
| | *Precision* | **0,492** | 0,461 | 0,455 | 0,452 |
| | *Recall* | **0,479** | 0,403 | 0,408 | 0,395 |
| | *F1-Score* | **0,453** | 0,360 | 0,361 | 0,364 |
| | *MCC* | **0,113** | 0,067 | 0,065 | 0,058 |
| | *Custom Score* | **0,588** | 0,514 | 0,523 | 0,509 |

Table 8.3.: Evaluation scores for the inclusion of channel descriptions. Entries in bold indicate a row maximum

| | | Prompt Types | | | | |
|---|---|---|---|---|---|---|
| | | **Baseline** | *LLM Generated* | *Zero-Shot CoT* | *Zero-Shot CoT (short)* | *ToT* |
| **Call** | *Accuracy* | 0,597 | 0,599 | 0,602 | **0,630** | 0,584 |
| | *Precision* | 0,603 | 0,603 | 0,591 | **0,609** | 0,591 |
| | *Recall* | 0,597 | 0,599 | 0,602 | **0,630** | 0,584 |
| | *F1-Score* | 0,599 | 0,598 | 0,596 | **0,618** | 0,587 |
| | *MCC* | 0,116 | **0,133** | 0,090 | 0,132 | 0,086 |
| **Chat** | *Accuracy* | **0,469** | 0,427 | 0,449 | 0,449 | 0,455 |
| | *Precision* | 0,459 | 0,376 | 0,463 | **0,476** | 0,414 |
| | *Recall* | **0,469** | 0,427 | 0,449 | 0,449 | 0,455 |
| | *F1-Score* | 0,382 | 0,335 | 0,370 | 0,373 | **0,399** |
| | *MCC* | 0,054 | 0,041 | 0,034 | **0,056** | 0,015 |
| **Web** | *Accuracy* | 0,397 | 0,434 | **0,465** | 0,442 | 0,375 |
| | *Precision* | 0,444 | 0,437 | **0,459** | 0,453 | 0,454 |
| | *Recall* | 0,397 | 0,434 | **0,465** | 0,442 | 0,375 |
| | *F1-Score* | 0,403 | 0,435 | **0,459** | 0,446 | 0,376 |
| | *MCC* | 0,119 | 0,110 | **0,134** | 0,126 | 0,125 |
| **App** | *Accuracy* | 0,451 | 0,469 | **0,496** | 0,467 | 0,429 |
| | *Precision* | 0,461 | 0,501 | **0,520** | 0,487 | 0,424 |
| | *Recall* | 0,451 | 0,469 | **0,496** | 0,467 | 0,429 |
| | *F1-Score* | 0,427 | 0,475 | **0,504** | 0,474 | 0,398 |
| | *MCC* | 0,164 | 0,199 | **0,219** | 0,174 | 0,126 |
| **Across Channel** | *Accuracy* | 0,479 | 0,483 | **0,503** | 0,497 | 0,461 |
| | *Precision* | 0,492 | 0,479 | **0,508** | 0,506 | 0,471 |
| | *Recall* | 0,479 | 0,483 | **0,503** | 0,497 | 0,461 |
| | *F1-Score* | 0,453 | 0,461 | **0,482** | 0,478 | 0,440 |
| | *MCC* | 0,113 | 0,121 | 0,119 | **0,122** | 0,088 |
| | *Custom Score* | 0,588 | 0,612 | **0,623** | 0,611 | 0,558 |

Table 8.4.: Evaluation scores for alternative prompt types. Entries in bold indicate a row maximum

| | | **Promising Combinations** | | |
|---|---|---|---|---|
| | | *1-shot CoT* | *3-shot CoT* | *5-shot CoT* |
| **Across Channel** | *Accuracy* | **0,534** | 0,480 | 0,507 |
| | *Precision* | 0,521 | 0,522 | **0,528** |
| | *Recall* | **0,534** | 0,480 | 0,507 |
| | *F1-Score* | **0,503** | 0,472 | 0,487 |
| | *MCC* | **0,159** | 0,137 | 0,153 |
| | *Custom Score* | 0,636 | 0,624 | **0,637** |

Table 8.5.: Evaluation scores for promising prompt combinations averaged across channels. Entries in bold indicate a row maximum

strictly better values while being a very close second with the custom score. Consequently, 1-shot CoT was chosen to be ran on the test set.

**Final results**

The chosen prompt (1-shot CoT) was then run on the test set once to exclude any bias. Table 8.6 details the results. To set the performance into perspective, three learning-learning based approaches were trained on the training set and evaluated on the test set: logistic regression[3], support vector classifier[4] and random forest[5]. We can observe that the random forest outperforms the LLM for each metric (across channels) but that for example the LLM scored higher values for the MCC than the regression and support vector classifier (across channels). This leaves the performance of the LLM-based classifier (given only one example through 1-shot) roughly in one range with the approaches trained on the entire data set.

Tables 8.1 - 8.6 showcase the variability of performance across the channels for each metric. Hence, the evaluation of the test run results calls for a detailed break-down of the classification per channel. For this purpose, the confusion matrices (CM) for each channel of the 1-shot CoT LLM run on the test set are shown in 8.1. A perfect classifier would have a diagonal matrix as confusion matrix which showcases the weakness of the given solution since there are many non-zero and far-from-zero entries at indexes away from the diagonal. The CM for the call channel shows that of the cases deemed "preferred" by the user or the LLM the majority are

[3]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, Last accessed 14/02/2025. Standard parameters taken except: *max_iter* which was set to 10000

[4]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC, Last accessed 14/02/2025. Standard parameters taken except *kernel*=linear, *probability*=True, *random_state*=0

[5]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, Last accessed 14/02/2025. Standard parameters taken

| | | | Learning Based Approaches | | |
|---|---|---|---|---|---|
| | | **1-shot CoT** | *Logistic Regression* | *Support Vector Classifier* | *Random Forest* |
| **Call** | *Accuracy* | 0,662 | 0,714 | **0,789** | 0,763 |
| | *Precision* | 0,684 | 0,673 | 0,622 | **0,691** |
| | *Recall* | 0,662 | 0,714 | **0,789** | 0,763 |
| | *F1-Score* | 0,672 | 0,693 | 0,696 | **0,720** |
| | *MCC* | 0,093 | 0,073 | 0,000 | **0,127** |
| **Chat** | *Accuracy* | 0,481 | 0,455 | 0,455 | **0,494** |
| | *Precision* | 0,451 | 0,425 | 0,379 | **0,479** |
| | *Recall* | 0,481 | 0,455 | 0,455 | **0,494** |
| | *F1-Score* | 0,412 | 0,430 | 0,412 | **0,460** |
| | *MCC* | 0,069 | 0,039 | 0,014 | **0,091** |
| **Web** | *Accuracy* | 0,471 | 0,494 | **0,506** | 0,503 |
| | *Precision* | 0,479 | 0,457 | 0,409 | **0,493** |
| | *Recall* | 0,471 | 0,494 | **0,506** | 0,503 |
| | *F1-Score* | 0,470 | 0,462 | 0,453 | **0,478** |
| | *MCC* | 0,154 | 0,137 | 0,148 | **0,154** |
| **App** | *Accuracy* | 0,510 | 0,568 | 0,558 | **0,571** |
| | *Precision* | 0,522 | 0,566 | 0,534 | **0,567** |
| | *Recall* | 0,510 | 0,568 | 0,558 | **0,571** |
| | *F1-Score* | 0,497 | 0,567 | 0,542 | **0,567** |
| | *MCC* | 0,245 | **0,302** | 0,264 | 0,296 |
| **Across Channel** | *Accuracy* | 0,531 | 0,558 | 0,577 | **0,583** |
| | *Precision* | 0,534 | 0,531 | 0,486 | **0,557** |
| | *Recall* | 0,531 | 0,558 | 0,577 | **0,583** |
| | *F1-Score* | 0,513 | 0,538 | 0,526 | **0,556** |
| | *MCC* | 0,140 | 0,138 | 0,106 | **0,167** |
| | *Custom Score* | 0,644 | 0,659 | 0,670 | **0,677** |

Table 8.6.: Evaluation scores for the comparison of the LLM approach with learning-based approaches on the test set. Entries in bold indicate a row maximum

actually “preferred”. On the other side, almost all cases where the customer labeled “would not use” (”no usage” in the figure for better visibility) were wrongly classified as “preferred" and not one case the LLM labeled “would not use” was correct. For the chat channel we see a similarly biased picture albeit the only label that is mostly predicted correctly is “acceptable”. If the customer labeled the chat “would not use”, the LLM is very likely to misclassify as “acceptable” which happened in 83% of the cases. App and web channel paint a slightly more balanced picture. Albeit still misclassifying more often than not - the “would not use” discrimination in the web channel works significantly better than for the call and chat channel while providing the correct predictions for the customer labels “preferred” and “acceptable” more often than any wrong ones. Interestingly, the second most frequent prediction for either of these labels is the respective other (which increases the agree-ability of the prediction as these are more similar than any combination with “would not use”)

## 8.2. Evaluation of Reasons Predictions

### 8.2.1. Methodology

#### Data

Data set splitting to prevent bias is not necessary for this analysis since there is no tuning or training. This means the entire filtered data set (few-shot examples, train and test set (see Section 8.1.1)) totaling 1556 data points was used to report the following results.

#### Metric

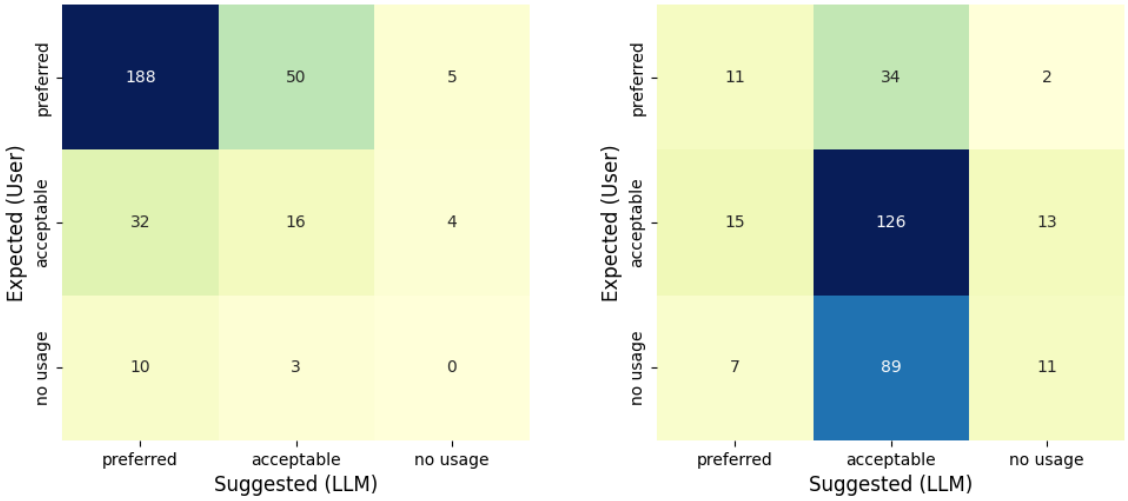
To enable the comparison of the provided LLM reasons with the original justification of the user, the strings were first embedded using Azure OpenAIs embedding model text-embedding-3-large. Subsequently, the similarity between the reasons generated by each approach and the original user reason was measured using the cosine metric as follows for embedding vectors a and b[6]:

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

### 8.2.2. Results

Table 8.7 shows the cosine similarity scores from the reasons presented by the user compared to those provided by the LLM. For each channel and preference, we can observe the mean similarity score for several comparisons. **User vs. LLM for User** describes the difference between the actual reason by the user and the one given by the LLM when asked why the user chose as she did. For the second approach – letting the LLM decide the channel preferences first and ask for its reasoning – there is a case distinction for the correctness of the decision of

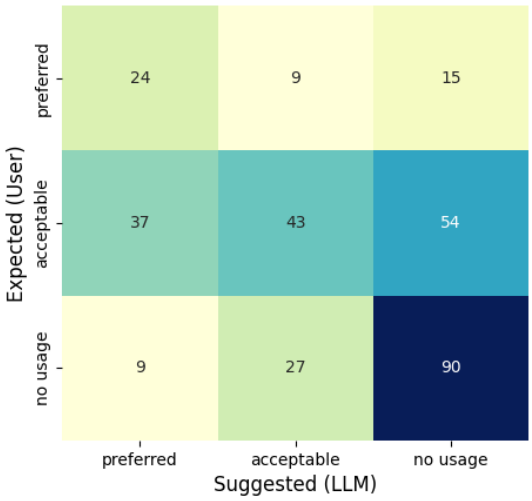
[6]As on: https://developers.google.com/machine-learning/clustering/dnn-clustering/supervised-similarity, Last accessed, 23/02/2025



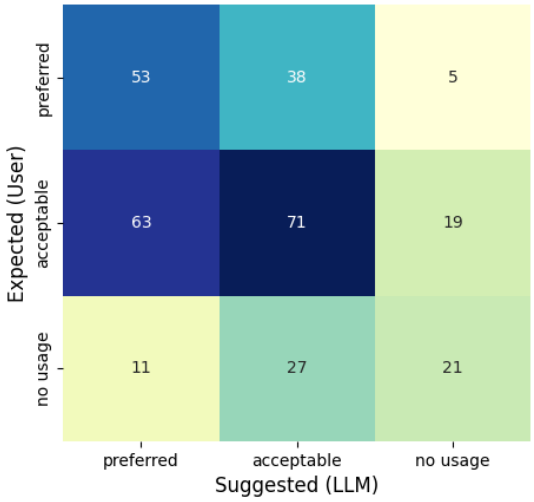

(a) Call




(b) Chat




(c) Self-Service App




(d) Self-Service Web

Figure 8.1.: Confusion matrices for 1-shot CoT performance on test data set. Labels shortened for visual reasons

the LLM (compared to the decision the user made). The last row specifies the mean values of the respective similarity scores across all the underlying values. These values might differ from the average obtained of the shown similarities per group due to different group sizes. The can see that the group-wise similarities range from 0.318 to 0.463 and general differences between the magnitudes of the scores between the groups. The reasons the LLM gave for the label "would not use" for the web channel in either approach seem to be less close to the original reasons than those given for the label "acceptable" for the call channel likely hinting at either more uniform reasons by the user in the latter and/or more commonly known phrases. We additionally observe that the average degree of similarity for the reasons

provided by the LLM from two perspectives: for the decision of the user and for the decision the LLM correctly made are almost identical and both larger than when the LLM provided a reason for its own wrong decision.

Looking at examples from the ten largest similarities for the reasons the LLM gave for the users' decision, we can see similarity scores between 0.72 and 0.76. In the following we want to give an idea of what the shared topics between the LLM and the user look like in the ten most similar examples. The most frequently represented combination under these is the chat channel and the label "would not use". The overlapping themes for this combination are the preference of speaking personally with an employee (assumingly via the phone) and that the chat takes too much time. For the reasons why the call is preferred, the overlapping topics also include the wish to speak to someone directly and the benefit of instant responses. The reasons provided by both the LLM and the user for why the app would not be used evolve around a general dislike of app. The chat is considered as preferred because of its simple convenience and the web would not be used because both the user and LLM mention to rather speak to someone personally.

| **Channel** | **Preference** | **User vs. LLM for User** | **User vs. LLM (correct decision)** | **User vs. LLM (wrong decision)** |
|---|---|---|---|---|
| Call | acceptable | 0,439 | 0,414 | 0,456 |
| Call | preferred | 0,452 | 0,460 | 0,423 |
| Call | would not use | 0,388 | 0,463 | 0,376 |
| Chat | acceptable | 0,361 | 0,354 | 0,364 |
| Chat | preferred | 0,418 | 0,427 | 0,403 |
| Chat | would not use | 0,417 | 0,406 | 0,361 |
| App | acceptable | 0,387 | 0,350 | 0,392 |
| App | preferred | 0,424 | 0,428 | 0,336 |
| App | would not use | 0,423 | 0,417 | 0,353 |
| Web | acceptable | 0,374 | 0,322 | 0,379 |
| Web | preferred | 0,403 | 0,407 | 0,352 |
| Web | would not use | 0,397 | 0,389 | 0,318 |
| *Mean across all underlying values* | | **0,417** | **0,417** | **0,368** |

Table 8.7.: Averaged similarity scores for the reasons that the LLM gave for the users' decision and those it gave for its own decision

# 9. Discussion

## 9.1. Interpretation of Results

This thesis provided insights into three topics: channel decision factors, LLM-powered channel preference prediction and generation of reasons for such preferences. The literature research and workshop/interviews with the case study company allowed for a more focused and updated overview of potentially relevant decision factors for customer channel choice while acknowledging the reality of the case study company with extensive experience in this field. This provides a basis for further data collection as it shows a wide spectrum of information that might be relevant for customers in making their decision. A first step was conducted with the case study company by conducting a survey for obtaining a data set for this specific use case focused on a subset of the previously identified factors.

The application of Large Language Models in predicting a customer service channel seems to be unprecedented but provided a first classifier without having to possess a large amount of training data. The development of the LLM-agent led to the insight that no single previously identified factor was the alone contributor to the performance of the model when specified in the model input but the combination of all of them together helped the most. As for the different prompting strategies, the inclusion of few-shots and alternative prompt types such as Chain-of-Thought prompting improved the performance compared to a simple hand-crafted prompt the opposite was true for the inclusion of channel descriptions provided by company experts. While lacking an approach currently in production to compare the developed solution with, a set of learning-based approaches such as a random forest provided similar but slightly better results. This shows how powerful LLMs can be without large amounts of training data. Nevertheless, on the one hand we could see strong differences in the performance across the different channels and on the other hand an overall result which does not exhibit the characteristics of a very good classifier. The final test set performance of around 0.64 for the custom score corresponds to an average of slightly more than 2.5 out of 4 points. This means if one label is predicted entirely wrong (considering a preferred choice as "would not use" resulting in 0 points), at least two of the remaining channels will need to have the correct label and one will have a somewhat agreeable label (e.g. acceptable instead of preferred). Despite the drawbacks, the LLM approach provides a portable solution since almost no preparation in terms of data collection and cleaning is necessary and thus works almost out of the box.

Reflecting on the approach for reason prediction capability, we tried masking different parts of information and let the LLM fill out the "gaps". The first approach meant providing the LLM with the user information and the decision made by the user and only ask for the likely

reason out of the perspective of the user. In the second approach we did not provide the user decision but asked the LLM to make this prediction based on the user information alone. We then inquired about the reasoning behind the LLM's decision which provided another reason to compare against the original user reason. Albeit scoring cosine similarities of over 0.7 for the most similar examples and catching common themes like the dislike of apps or the preference of direct personal contact, the average similarity score revolves around only 0.41. As was to be expected, the similarity of the reason for the decision the LLM made was notably lower in case a different decision was made by the user.

## 9.2. Managerial Implications

The findings of this thesis point to several managerial implications. Effective data collection is a crucial process, as it provides a comprehensive overview of customer behavior and uncovers opportunities for developing innovative and data driven applications. The results from the LLM-input experiments suggests that channel preference is best identified by looking at a holistic picture instead of just a few individual factors, indicating a need for the structured availability of as much information as possible about the customer in the live systems. To bridge potential data gaps, managers could explore the use of LLMs for decision making in data scarce environments until sufficient high-quality data is collected for the use of (cheaper and faster) traditional ML-based solutions. Ultimately, the development of an omnichannel strategy would allow in perspective to switch between channels seamlessly creating a truly integrated customer experience. This requires the build-up of an overarching data infrastructure independent of the individual channels.

## 9.3. Limitations

The limitations of the structured literature review for the channel choice determinants revolve around the fact that only one researcher conducted the screening of articles and synthesis of factors. The significance and spectrum of the answers in the workshop and interviews is limited to the participants expertise. The data set obtained through the evaluation is only as realistic as it can be without collecting live data from the channels. Furthermore, the fact that each participant answered three scenarios introduces less diversity in the answers. This was tried to be countered to some degree by splitting the train and test set on the persons instead of individual answers. Finally, the use of only one LLM by one provider limits the generalizability.

## 9.4. Future Work

Future research directions for the channel choice determinants (factors) could include the expansion to other application areas and companies to achieve a more comprehensive understanding of channel choice and potentially switching behavior. The approach of this

project could be replicated and verified with a variety of different LLM models and providers, especially considering the fast-paced and competitive field of LLM development. Another interesting approach could be to fine-tune and evaluate LLMs on other use case specific data. With respect to the overall problem of channel preference prediction, specialized machine-learning solutions could be developed and fine-tuned based on similar data to achieve potentially higher classification scores. In terms of improving the data quality, a system for collecting channel preferences from incoming customer requests could be implemented into a production system. When thinking about omnichannel as a goal in customer service, an interesting question would be how to enable a fluent transition between channels by continuously collecting all mentioned data points from the customer independent of the channel and providing it to the target one.

# 10. Conclusion

This thesis was concerned with the use of Large Language Models in the realm of customer service and communication channel choice specifically. Guided by three research questions, the first step was to generate an understanding of customer channel choice determinants by the means of a structured literature review and a workshop and interviews with a large European Insurance company to answer the question of what relevant factors for deciding the optimal channel for customer service requests are. This yielded almost 70 factors to be of interest in making a choice for a communication channel. Building on this, a comprehensive survey in cooperation with the case study company generated the first comprehensive data set for channel switching preferences in the company. To address the second research question of how different input factors and prompt strategies are influencing the effectiveness of LLMs in selecting appropriate communication channels for customer service requests, an LLM-based navigation agent was built to determine the channel preferences for incoming requests and different input and prompt strategies were tested for their influence on performance. In terms of input to the LLM agent, the results indicated that the inclusion of no single previously obtained factor individually had an outstanding impact on performance. The best option here was to include all the factors together. For the different prompt strategies, few-shots and Chain-of-Thought type of prompts yielded stronger results than a simple handwritten prompt. The inclusion of detailed channel descriptions collected from channel experts within the case study company even had a negative impact on model performance. Compared with learning-based approaches such as logistic regression, support vector classifiers or random forests the LLM-approach was able to perform competitively in predicting customer channel preferences. In the light of the third research question – how well do LLMs predict the reasons for choosing a customer service channel? – the capability of the LLM to produce a reason for the chosen channel was analyzed. Two different approaches were tested to obtain a reason like the one given by the user: one by asking the LLM to explain the user decision and another by letting the LLM decide on its' own and explain this decision afterwards. When the LLM decision matched the one by the user, both approaches yielded almost identical similarity scores compared with the original users' answer.

# A. Scenarios

The scenarios presented to users in the survey were written in German. Each one starts with a request to put yourself in the situation of having called the hotline but not yet being connected to a human employee (i.e. being in the queue). The following shows the English translation of a liability situation (low complexity and new claim):

"Please imagine the following situation: You call the customer service of your insurance company, but are not yet connected to an employee. You have accidentally dropped a friend's cell phone and it has been damaged. You want to report this simple liability claim to your insurance company. You have not yet contacted the insurance company to report this claim"

The following tables are showing all the used scenarios in their original wording:

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen komplizierten Schadenfall gemeldet. Sie haben einen Schlüssel der zentralen Schließanlage Ihres Büros verloren. Der Verlust zieht sehr hohe Kosten mit sich. Jetzt wollen Sie sich bei Ihrem Versicherungsunternehmen über den aktuellen Bearbeitungsstand erkundigen. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen simplen Schaden gemeldet. Sie haben letzte Woche den Wohnungstürschlüssel in Ihrer Wohnung vergessen und haben sich ausgesperrt. Deshalb mussten Sie einen Schlüsseldienst beauftragen. Jetzt wollen Sie sich bei Ihrem Versicherungsunternehmen über den aktuellen Bearbeitungsstand erkundigen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadensabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben einen Schlüssel der zentralen Schließanlage in ihrem Büro verloren. Der Verlust zieht sehr hohe Kosten mit sich. Diesen komplizierten Schaden wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben letzte Woche den Wohnungstürschlüssel in Ihrer Wohnung vergessen und haben sich ausgesperrt. Deshalb mussten Sie einen Schlüsseldienst beauftragen. Diesen Schaden wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table A.1.: Exemplary scenarios for the intent: lost keys

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen komplizierten Leitungsschaden gemeldet. Dabei handelt es sich um einen undichten Warmwasserboiler, der zu großflächigen Feuchtigkeitsschäden in mehreren Räumen geführt hat. Jetzt wollen Sie sich bei Ihrer Versicherung über den aktuellen Bearbeitungsstand erkundigen und wann dieser Schaden final von einer Fachfirma behoben sein wird. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen einfachen Wasserschaden gemeldet. Dabei handelt es sich um einen einfachen Wasserschaden durch eine Wasserleitung unter der Spüle in Ihrer Küche. Jetzt wollen Sie sich bei Ihrer Versicherung über den aktuellen Bearbeitungsstand erkundigen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadensabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. In Ihrem Neubau ist anscheinend der Warmwasserboiler undicht und hat in mehreren Räumen zu großflächigen Feuchtigkeitsschäden geführt. Diesen umfangreichen Leitungsschaden wollen Sie Ihrem Versicherungsunternehmen melden, deswegen rufen Sie an. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen den Kundenservice Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben aktuell einen unkomplizierten Wasserschaden durch eine Wasserleitung unter der Spüle in Ihrer Küche. Diesen unkomplizierten Wasserschaden wollen Sie Ihrem Versicherungsunternehmen melden, deswegen rufen Sie an. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table A.2.: Exemplary scenarios for the intent: pipe damage

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben Ihrer Versicherung bereits vor einiger Zeit einen großen Einbruch gemeldet. Bei dem Einbruch wurden mehrere wertvolle Gegenstände und auch Ihr Auto gestohlen. Die Polizei war vor Ort und hat den Einbruch aufgenommen, aber sich bei Ihnen noch nicht wieder gemeldet. Jetzt wollen Sie sich über den Bearbeitungsstand bei Ihrer Versicherung erkundigen. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben Ihrer Versicherung bereits vor einiger Zeit einen simplen Einbruch gemeldet. Bei Ihnen wurde vor einer Woche in Ihr Haus eingebrochen. Dabei wurde lediglich ein Fernseher entwendet. Die Polizei hat den Einbruch bereits aufgenommen. Jetzt wollen Sie sich über den Bearbeitungsstand bei Ihrer Versicherung erkundigen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadensabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Heute Nacht wurde in Ihr Haus eingebrochen. Bei dem Einbruch wurden mehrere wertvolle Gegenstände und auch Ihr Auto gestohlen. Die Polizei ist bereits vor Ort, aber untersucht noch den Tatort. Diesen großen Einbruch wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen den Kundenservice Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Bei Ihnen wurde vor einer Woche in die Wohnung eingebrochen. Dabei wurde lediglich ein Fernseher entwendet. Die Polizei hat den Einbruch bereits aufgenommen. Diesen aus ihrer Sicht simplen Fall wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table A.3.: Exemplary scenarios for the intent: break-in

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen komplizierten Haftpflichtschaden gemeldet. Ihr Sohn ist beim Fahrradfahren gestürzt und hat dabei mehrere teure Autos zerkratzt. Jetzt wollen Sie sich bei Ihrer Versicherung über den Bearbeitungsstand und die weitere Vorgehensweise erkundigen. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen Haftpflichtschaden gemeldet. Sie hatten versehentlich das Handy Ihres Freundes fallenlassen, dabei ist es beschädigt worden. Jetzt wollen Sie sich über den Bearbeitungsstand bei Ihrer Versicherung erkundigen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenshotline Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Ihre Tochter ist beim Fahrradfahren gestürzt und hat dabei mehrere teure Autos zerkratzt. Diesen komplizierten Haftpflichtversicherungsfall wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen den Kundenservice Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben versehentlich das Handy von einem Freund fallengelassen, dabei wurde es beschädigt. Diesen simplen Haftpflichtschaden wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table A.4.: Exemplary scenarios for the intent: liability

| Scenario combinations | | |
|---|---|---|
| | **High complexity** | **Low complexity** |
| **Existing claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen komplizierten Unfallschaden gemeldet. Sie sind mit dem Auto gefahren und dabei mit einem Fahrrad kollidiert. Dabei wurde der Fahrradfahrer verletzt. Mittlerweile hat sich herausgestellt, dass bei dem Unfall noch eine weitere Person zu Schaden gekommen ist. Jetzt wollen Sie sich bei Ihrer Versicherung über den aktuellen Stand erkundigen und dem Versicherungsunternehmen zusätzlich melden, dass eine weitere Person zu Schaden gekommen ist. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben der Versicherung bereits vor einiger Zeit einen simplen Park-Unfall gemeldet. Sie haben beim Parken ein anderes Auto touchiert. Dabei ist es zu einem Blechschaden gekommen. An dem Unfall war außer Ihnen niemand beteiligt. Jetzt wollen Sie sich bei Ihrer Versicherung über den Bearbeitungsstand erkundigen und die nächsten Schritte besprechen. |
| **New claim** | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadensabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie sind gerade eben als Autofahrer mit einem Fahrrad kollidiert. Dabei wurde der Fahrradfahrer verletzt. Diesen komplizierten Unfall wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. | Bitte stellen Sie sich folgende Situation vor: Sie rufen die Schadenabteilung Ihres Versicherungsunternehmens an, sind aber noch nicht mit einem Mitarbeiter verbunden. Sie haben beim Parken ein anderes Auto touchiert. Dabei ist es zu einem Blechschaden ohne Fremdbeteiligung gekommen. Diesen simplen Park-Unfall wollen Sie Ihrem Versicherungsunternehmen melden. Bisher hatten Sie die Versicherung noch nicht kontaktiert, um diesen Schaden zu melden. |

Table A.5.: Exemplary scenarios for the intent: accident

# B. Prompts

**Baseline system instruction prompt**: *You are an assistant for a customer service center of an insurance company. Your goal is to help customers navigate to the appropriate channel based on a variety of factors which are given in the user prompt. The available channels are: Call, Chat, Self-Service Web, Self-Service App. Assign a label to each channel. The labels are: preferred, acceptable, undesired. Multiple channels with the same label are possible. Provide your answer as a JSON string with the channel names as keys.*

**LLM generated prompt**: *Assign labels to the available customer service channels based on the factors provided by the user. Consider the factors provided by the user to determine the suitability of each channel. The available channels are: Call, Chat, Self-Service Web, Self-Service App. The labels to assign are: preferred, acceptable, undesired. Multiple channels can have the same label.*
*# Steps*
*1. Analyze the factors provided by the user.*
*2. Determine the suitability of each channel based on these factors.*
*3. Assign the appropriate label to each channel.*
*4. Ensure that the output is formatted as a JSON string with the channel names as keys and their labels as values.*
*# Output format*
*Provide the output as a JSON string with the channel names as keys and their labels as values. Example format:*
*{*
*"Call": [label],*
*"Chat": [label],*
*"Self-Service Web": [label],*
*"Self-Service App": [label],*
*}*

**Zero-shot Chain-of-Thought**: *You are an assistant for a customer service center of an insurance company. Your goal is to help customers navigate to the appropriate channel based on a variety of factors which are given in the user prompt. The available channels are: Call, Chat, Self-Service Web, Self-Service App. Assign a label to each channel. The labels are: preferred, acceptable, undesired. Multiple channels with the same label are possible. Provide your answer as a JSON string with the channel names as keys. Before that, list all of your steps explicitly.*
*Let's think step by step.*

*1. First let's understand who the user is. This involves looking at the age and innovativeness.*

*2. Then, we take a look at the users' habits: the previously used channels for communicating with the company and the devices he or she ownes.*
*3. In a last step we take into consideration what situation the user is in at the moment. This involves the intent and complexity of the matter they are contacting about in addition to whether they did already contact the company about this claim already.*
*4. We should also think about what the customer said in plain text to make our final analysis.*

**Zero-shot Chain-of-Thought (short)**: *You are an assistant for a customer service center of an insurance company. Your goal is to help customers navigate to the appropriate channel based on a variety of factors which are given in the user prompt. The available channels are: Call, Chat, Self-Service Web, Self-Service App. Assign a label to each channel. The labels are: preferred, acceptable, undesired. Multiple channels with the same label are possible. Provide your answer as a JSON string with the channel names as key. Before that, list all of your steps explicitly. Let's think step by step.*

**Chain-of-Thought**: *You are an assistant for a customer service center of an insurance company. Your goal is to help customers navigate to the appropriate channel based on a variety of factors which are given in the user prompt. The available channels are: Call, Chat, Self-Service Web, Self-Service App. Assign a label to each channel. The labels are: preferred, acceptable, undesired. Multiple channels with the same label are possible. Provide your answer as a JSON string with the channel names as keys. Before that, list all of your steps explicitly.*
*The following shows some examples:*

*Exemplary user prompt: <User Input>*
*Let's think step by step.*

*1. First let's understand who the user is. This involves looking at the age and innovativeness.*
*2. Then, we take a look at the users' habits: the previously used channels for communicating with the company and the devices he or she ownes.*
*3. In a last step we take into consideration what situation the user is in at the moment. This involves the intent and complexity of the matter they are contacting about in addition to whether they did already contact the company about this claim already.*
*4. We should also think about what the customer said in plain text to make our final analysis.*

*The expected answer is:*
*{*
*"Call": [label],*
*"Chat": [label],*
*"Self-Service Web": [label],*
*"Self-Service App": [label],*
*}*

**Tree-of-Thought (short)**: *Imagine three different experts are working on a problem. The problem is*

*to decide which channel is the best choice for a customer of an insurance company. All experts will write down 1 step of their thinking based on the information given in the input, then share it with the group. Then all experts will go on to the next step, etc. If any expert realises they're wrong at any point then they leave. The available channels are: Call, Chat, Self-Service Web, Self-Service App. Assign a label to each channel. The labels are: preferred, acceptable, undesired. Multiple channels with the same label are possible. Provide your answer as a JSON string with the channel names as keys. Before that, list all of your steps explicitly.*

# C. Code Snippets

Listing C.1: Python code for calling the Azure OpenAI API

```python
def get_channel_categorization_generic(
        client, # e.g.: AzureOpenAI
        model="gpt-4o",
        system_prompt="",
        user_prompt=""
    ) -> tuple[MultiChannelCategorization, str]:
        response = client.chat.completions.create(
            model=model,
            messages=[
                {
                "role": "system",
                "content": system_prompt
                },
                {
                "role": "user",
                "content": user_prompt
                }
            ],
            temperature=0.00001,
            seed=42,
            max_tokens=100
        )
        raw_response = response.choices[0].message.content

        json_string = extract_JSON(input_string=raw_response)
        response_dict = parse_keys_values(json_string, english_labels=True)

        return MultiChannelCategorization(
            channels=[ChannelCategorization(
                channel_name=[_channel.value[0] for _channel in Channels if
                    _channel.value[0] == str(channel_key_name)][0],
                channel_category=response_dict[channel_key_name]
            ) for channel_key_name in list(response_dict.keys())]
        ), raw_response
```

```
34
35 class ChannelCategorization(BaseModel):
36     channel_name: str
37     channel_category: str
38     category_reason: str = ""
39
40
41 class MultiChannelCategorization(BaseModel):
42     channels: list[ChannelCategorization]
```

# List of Figures

# List of Tables

# Bibliography


[1] Salesforce. *27 famous quotes about customer service from CEOs & business leaders - Salesforce.com*. en. n.d. URL: https://www.salesforce.com/ca/hub/service/famous-customer-service-quotes/ (visited on 02/07/2025).

[2] M. M. Hasan, S. A. Jalal Siam, and A. Haque. "THE SIGNIFICANCE OF CUSTOMER SERVICE IN ESTABLISHING TRUST AND ENHANCING THE REPUTATION OF THE BANKING INDUSTRY IN BANGLADESH". en. In: *Business and Economics in Developing Countries* 1.2 (May 2023), pp. 71–75. ISSN: 29909449. DOI: 10.26480/bedc.02.2023.71.75. URL: https://bedc.com.my/paper/2bedc2023/2bedc2023-71-75.pdf.

[3] Y. Skaf, C. Eid, A. Thrassou, S. E. Nemar, and K. S. Rebeiz. "Technology and service quality: achieving insurance industry customer satisfaction and loyalty under crisis conditions". en. In: *EuroMed Journal of Business* ahead-of-print.ahead-of-print (Sept. 2024). Publisher: Emerald Publishing Limited. ISSN: 1450-2194. DOI: 10.1108/EMJB-01-2024-0027. URL: https://www.emerald.com/insight/content/doi/10.1108/emjb-01-2024-0027/full/html.

[4] Z. Liu, C. Long, X. Lu, Z. Hu, J. Zhang, and Y. Wang. "Which Channel to Ask My Question?: Personalized Customer Service Request Stream Routing Using Deep Reinforcement Learning". In: *IEEE Access* 7 (2019). Conference Name: IEEE Access, pp. 107744–107756. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2932047. URL: https://ieeexplore.ieee.org/abstract/document/8784156.

[5] A. Iacoviello and A. Downie. *What Is Customer Care? | IBM*. en. Apr. 2024. URL: https://www.ibm.com/topics/customer-care (visited on 12/12/2024).

[6] S. K. Roy, W. M. Lassar, S. Ganguli, B. Nguyen, and X. Yu. "Measuring service quality: a systematic review of literature". en. In: *International Journal of Services, Economics and Management* 7.1 (2015), p. 24. ISSN: 1753-0822, 1753-0830. DOI: 10.1504/IJSEM.2015.076322. URL: http://www.inderscience.com/link.php?id=76322.

[7] J. Majava and V. Isoherranen. "Business model evolution of customer care services". en. In: *Journal of Industrial Engineering and Management* 12.1 (Jan. 2019), p. 1. ISSN: 2013-0953, 2013-8423. DOI: 10.3926/jiem.2725. URL: http://www.jiem.org/index.php/jiem/article/view/2725.

[8] P. C. Verhoef, P. K. Kannan, and J. J. Inman. "From Multi-Channel Retailing to Omni-Channel Retailing: Introduction to the Special Issue on Multi-Channel Retailing". In: *Journal of Retailing*. Multi-Channel Retailing 91.2 (June 2015), pp. 174–181. ISSN: 0022-4359. DOI: 10.1016/j.jretai.2015.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0022435915000214.

[9] X.-L. Shen, Y.-J. Li, Y. Sun, and N. Wang. "Channel integration quality, perceived fluency and omnichannel service usage: The moderating roles of internal and external usage experience". In: *Decision Support Systems* 109 (2018), pp. 61–73. DOI: 10.1016/j.dss.2018.01.006. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041597744&doi=10.1016%2fj.dss.2018.01.006&partnerID=40&md5=b6f0c8bade05a248efe202489da6be84.

[10] C. Lazaris and A. Vrechopoulos. *From Multichannel to "Omnichannel" Retailing: Review of the Literature and Calls for Research*. June 2014. DOI: 10.13140/2.1.1802.4967.

[11] C. Gerea, F. Gonzalez-Lopez, and V. Herskovic. "Omnichannel Customer Experience and Management: An Integrative Review and Research Agenda". en. In: *Sustainability* 13.5 (Jan. 2021). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 2824. ISSN: 2071-1050. DOI: 10.3390/su13052824. URL: https://www.mdpi.com/2071-1050/13/5/2824.

[12] T. M. T. Hossain, S. Akter, U. Kattiyapornpong, and Y. K. Dwivedi. "Multichannel integration quality: A systematic review and agenda for future research". In: *Journal of Retailing and Consumer Services* 49 (July 2019), pp. 154–163. ISSN: 0969-6989. DOI: 10.1016/j.jretconser.2019.03.019. URL: https://www.sciencedirect.com/science/article/pii/S0969698919300700.

[13] J. Reardon and D. E. McCorkle. "A consumer model for channel switching behavior". In: *International Journal of Retail & Distribution Management* 30.4 (Jan. 2002). Publisher: MCB UP Ltd, pp. 179–185. ISSN: 0959-0552. DOI: 10.1108/09590550210423654. URL: https://doi.org/10.1108/09590550210423654.

[14] S. Yan, T. W. Archibald, X. Han, and Y. Bian. "Whether to adopt "buy online and return to store" strategy in a competitive market?" en. In: *European Journal of Operational Research* 301.3 (Sept. 2022), pp. 974–986. ISSN: 03772217. DOI: 10.1016/j.ejor.2021.11.040. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221721009917.

[15] Z. Li, W. Yang, and X. Chen. "Omnichannel inventory models accounting for Buy-Online–Return-to-Store service and random demand". en. In: *Soft Computing* 25.17 (Sept. 2021), pp. 11691–11710. ISSN: 1433-7479. DOI: 10.1007/s00500-021-06045-0. URL: https://doi.org/10.1007/s00500-021-06045-0.

[16] M.-J. Miquel-Romero, M. Frasquet, and A. Molla-Descals. "The role of the store in managing postpurchase complaints for omnichannel shoppers". In: *Journal of Business Research* 109 (2020), pp. 288–296. DOI: 10.1016/j.jbusres.2019.09.057. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076724603&doi=10.1016%2fj.jbusres.2019.09.057&partnerID=40&md5=3f4406c77ce762ced23dd3ed49fa8d4f.

[17] L. Wolf and M. Steul-Fischer. "Factors of customers' channel choice in an omnichannel environment: a systematic literature review". en. In: *Management Review Quarterly* 73.4 (Dec. 2023), pp. 1579–1630. ISSN: 2198-1639. DOI: 10.1007/s11301-022-00281-w. URL: https://doi.org/10.1007/s11301-022-00281-w.

[18] R. T. Rust and M.-H. Huang. "The Service Revolution and the Transformation of Marketing Science." eng. In: *Marketing Science* 33.2 (Mar. 2014). Publisher: INFORMS: Institute for Operations Research, pp. 206–221. ISSN: 0732-2399. DOI: 10.1287/mksc.2013.0836. URL: https://research.ebsco.com/linkprocessor/plink?id=e949df7a-2272-3295-b5de-d94652e890f1.

[19] M. Adam, M. Wessel, and A. Benlian. "AI-based chatbots in customer service and their effects on user compliance". en. In: *Electronic Markets* 31.2 (June 2021), pp. 427–445. ISSN: 1422-8890. DOI: 10.1007/s12525-020-00414-7. URL: https://doi.org/10.1007/s12525-020-00414-7.

[20] J. Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". en. In: *Communications of the ACM* 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/365153.365168. URL: https://dl.acm.org/doi/10.1145/365153.365168.

[21] E. Adamopoulou and L. Moussiades. "An Overview of Chatbot Technology". en. In: *Artificial Intelligence Applications and Innovations*. Ed. by I. Maglogiannis, L. Iliadis, and E. Pimenidis. Vol. 584. Series Title: IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2020, pp. 373–383. ISBN: 978-3-030-49185-7 978-3-030-49186-4. DOI: 10.1007/978-3-030-49186-4_31. URL: https://link.springer.com/10.1007/978-3-030-49186-4_31.

[22] N. Pfeuffer, A. Benlian, H. Gimpel, and O. Hinz. "Anthropomorphic Information Systems". en. In: *Business & Information Systems Engineering* 61.4 (Aug. 2019), pp. 523–533. ISSN: 2363-7005, 1867-0202. DOI: 10.1007/s12599-019-00599-y. URL: http://link.springer.com/10.1007/s12599-019-00599-y.

[23] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner. "AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives". en. In: *Business & Information Systems Engineering* 61.4 (Aug. 2019), pp. 535–544. ISSN: 2363-7005, 1867-0202. DOI: 10.1007/s12599-019-00600-8. URL: http://link.springer.com/10.1007/s12599-019-00600-8.

[24] A. Ramesh and V. Chawla. "Chatbots in Marketing: A Literature Review Using Morphological and Co-Occurrence Analyses". en. In: *Journal of Interactive Marketing* 57.3 (Aug. 2022). Publisher: SAGE Publications, pp. 472–496. ISSN: 1094-9968. DOI: 10.1177/10949968221095549. URL: https://doi.org/10.1177/10949968221095549.

[25] L. Wang, N. Huang, Y. Hong, L. Liu, X. Guo, and G. Chen. "Voice-based AI in call center customer service: A natural field experiment". en. In: *Production and Operations Management* 32.4 (Apr. 2023). Publisher: SAGE Publications, pp. 1002–1018. ISSN: 1059-1478. DOI: 10.1111/poms.13953. URL: https://doi.org/10.1111/poms.13953.

[26] G. G. Chowdhury. "Natural Language Processing". en. In: *Annual Review of Information Science and Technology* 37 (2003), pp. 51–89. ISSN: 0066-4200.

[27] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, and M. Vassilakopoulos. "Large Language Models versus Natural Language Understanding and Generation". en. In: *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*. Lamia Greece: ACM, Nov. 2023, pp. 278–290. ISBN: 9798400716263. DOI: 10.1145/3635059.3635104. URL: https://dl.acm.org/doi/10.1145/3635059.3635104.

[28] R. Shaik and K. S. Kishore. *Enhancing Text Generation in Joint NLG/NLU Learning Through Curriculum Learning, Semi-Supervised Training, and Advanced Optimization Techniques*. arXiv:2410.13498. Oct. 2024. URL: http://arxiv.org/abs/2410.13498.

[29] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. *A Survey of Large Language Models*. arXiv:2303.18223. Oct. 2024. URL: http://arxiv.org/abs/2303.18223.

[30] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, and D. Roth. *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*. arXiv:2111.01243. Nov. 2021. URL: http://arxiv.org/abs/2111.01243.

[31] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics,and Speech Recognition with Language Models*. 3rd. Aug. 2024. URL: https://web.stanford.edu/~jurafsky/slp3/ (visited on 11/30/2024).

[32] Y. Bengio, R. Ducharme, and P. Vincent. "A Neural Probabilistic Language Model". In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2000.

[33] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun. "Pre-Trained Language Models and Their Applications". In: *Engineering* 25 (June 2023), pp. 51–65. ISSN: 2095-8099. DOI: 10.1016/j.eng.2022.04.024. URL: https://www.sciencedirect.com/science/article/pii/S2095809922006324.

[34] S. Edunov, A. Baevski, and M. Auli. *Pre-trained Language Model Representations for Language Generation*. arXiv:1903.09722. Apr. 2019. URL: http://arxiv.org/abs/1903.09722.

[35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. "Improving Language Understanding by Generative Pre-Training". In: 2018. URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". en. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017).

[37] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. *Scaling Laws for Neural Language Models*. arXiv:2001.08361. Jan. 2020. URL: http://arxiv.org/abs/2001.08361.

[38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. *Language Models are Few-Shot Learners*. en. arXiv:2005.14165 [cs]. July 2020. URL: http://arxiv.org/abs/2005.14165.

[39] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. "PaLM: Scaling Language Modeling with Pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113. ISSN: 1533-7928. URL: http://jmlr.org/papers/v24/22-1144.html.

[40] GitHub. *GitHub Copilot · Your AI pair programmer*. en. n.d. URL: https://github.com/features/copilot (visited on 12/01/2024).

[41] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. *Evaluating Large Language Models Trained on Code*. arXiv:2107.03374. July 2021. URL: http://arxiv.org/abs/2107.03374.

[42] R. Koshkin, K. Sudoh, and S. Nakamura. *TransLLaMa: LLM-based Simultaneous Translation System*. arXiv:2402.04636. Feb. 2024. DOI: 10.48550/arXiv.2402.04636. URL: http://arxiv.org/abs/2402.04636.

[43] H. Xiong, J. Bian, Y. Li, X. Li, M. Du, S. Wang, D. Yin, and S. Helal. *When Search Engine Services meet Large Language Models: Visions and Challenges*. arXiv:2407.00128. June 2024. URL: http://arxiv.org/abs/2407.00128.

[44] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. en. July 2023. DOI: 10.36227/techrxiv.23589741.v1. URL: https://www.techrxiv.org/doi/full/10.36227/techrxiv.23589741.v1.

[45] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting. *Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do*. arXiv:2103.11790. Feb. 2022. URL: http://arxiv.org/abs/2103.11790.

[46] A. Caliskan, J. J. Bryson, and A. Narayanan. *Semantics derived automatically from language corpora contain human-like biases*. arXiv:1608.07187. May 2017. URL: http://arxiv.org/abs/1608.07187.

[47] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, and N. Muennighoff. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. en. 2023. URL: https://inria.hal.science/hal-03850124v1.

[48] BDEW. *Pro-Kopf-Stromverbrauch in Deutschland 2023*. de. Sept. 2024. URL: https://de.statista.com/statistik/daten/studie/240696/umfrage/pro-kopf-stromverbrauch-in-deutschland/ (visited on 01/25/2025).

[49] A. S. Luccioni, S. Viguier, and A.-L. Ligozat. *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*. arXiv:2211.02001. Nov. 2022. URL: http://arxiv.org/abs/2211.02001.

[50] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. "Explainability for Large Language Models: A Survey". en. In: *ACM Transactions on Intelligent Systems and Technology* 15.2 (Apr. 2024), pp. 1–38. ISSN: 2157-6904, 2157-6912. DOI: 10.1145/3639372. URL: https://dl.acm.org/doi/10.1145/3639372.

[51] F. Doshi-Velez and B. Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608. Mar. 2017. URL: http://arxiv.org/abs/1702.08608.

[52] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. *On Faithfulness and Factuality in Abstractive Summarization*. arXiv:2005.00661. May 2020. URL: http://arxiv.org/abs/2005.00661.

[53] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". en. In: *ACM Transactions on Information Systems* (Nov. 2024), p. 3703155. ISSN: 1046-8188, 1558-2868. DOI: 10.1145/3703155. URL: https://dl.acm.org/doi/10.1145/3703155.

[54] S. Parikh, Q. Vohra, P. Tumbade, and M. Tiwari. *Exploring Zero and Few-shot Techniques for Intent Classification*. arXiv:2305.07157. May 2023. URL: http://arxiv.org/abs/2305.07157.

[55] B. Lajčinová, P. Valábek, and M. Spišiak. *Intent Classification for Bank Chatbots through LLM Fine-Tuning*. arXiv:2410.04925. Oct. 2024. URL: http://arxiv.org/abs/2410.04925.

[56] X. Sun, X. Li, S. Zhang, S. Wang, F. Wu, J. Li, T. Zhang, and G. Wang. *Sentiment Analysis through LLM Negotiations*. arXiv:2311.01876. Nov. 2023. URL: http://arxiv.org/abs/2311.01876.

[57] A. Shahnaz Ipa, P. Nath Roy, M. Abu Tareq Rony, A. Raza, N. Latif Fitriyani, Y. Gu, and M. Syafrudin. "BdSentiLLM: A Novel LLM Approach to Sentiment Analysis of Product Reviews". In: *IEEE Access* 12 (2024). Conference Name: IEEE Access, pp. 189330–189343. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3516826. URL: https://ieeexplore.ieee.org/abstract/document/10798428.

[58] Z. Cheng, W. Zhang, C.-C. Chou, Y.-Y. Jau, A. Pathak, P. Gao, and U. Batur. "E-Commerce Product Categorization with LLM-based Dual-Expert Classification Paradigm". In: *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. Ed. by S. Kumar, V. Balachandran, C. Y. Park, W. Shi, S. A. Hayati, Y. Tsvetkov, N. Smith, H. Hajishirzi, D. Kang, and D. Jurgens. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 294–304. DOI: 10.18653/v1/2024.customnlp4u-1.22. URL: https://aclanthology.org/2024.customnlp4u-1.22/.

[59] A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi. *Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making*. en. Feb. 2023. DOI: 10.1101/2023.02.02.23285399. URL: http://medrxiv.org/lookup/doi/10.1101/2023.02.02.23285399.

[60] S. Fatemi, Y. Hu, and M. Mousavi. *A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification*. arXiv:2411.02476 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2411.02476. URL: http://arxiv.org/abs/2411.02476.

[61] N. Tyson and V. Matula. "Improved lsi-based natural language call routing using speech recognition confidence scores". In: *Second IEEE International Conference on Computational Cybernetics, 2004. ICCC 2004.* Aug. 2004, pp. 409–413. DOI: 10.1109/ICCCYB.2004.1437763. URL: https://ieeexplore.ieee.org/document/1437763.

[62] A. R. Hevner, S. T. March, J. Park, and S. Ram. "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1 (2004). Publisher: Management Information Systems Research Center, University of Minnesota, pp. 75–105. ISSN: 0276-7783. DOI: 10.2307/25148625. URL: https://www.jstor.org/stable/25148625.

[63] S. A. Neslin, D. Grewal, R. Leghorn, V. Shankar, M. L. Teerling, J. S. Thomas, and P. C. Verhoef. "Challenges and Opportunities in Multichannel Customer Management". en. In: *Journal of Service Research* 9.2 (Nov. 2006), pp. 95–112. ISSN: 1094-6705, 1552-7379. DOI: 10.1177/1094670506293559. URL: https://journals.sagepub.com/doi/10.1177/1094670506293559.

[64] C.-C. Hsiao, H. Ju Rebecca Yen, and E. Y. Li. "Exploring consumer value of multichannel shopping: a perspective of means-end theory". In: *Internet Research* 22.3 (Jan. 2012). Publisher: Emerald Group Publishing Limited, pp. 318–339. ISSN: 1066-2243. DOI: 10.1108/10662241211235671. URL: https://doi.org/10.1108/10662241211235671.

[65] M. Immonen, S. Sintonen, and J. Koivuniemi. "The value of human interaction in service channels". In: *Computers in Human Behavior* 78 (Jan. 2018), pp. 316–325. ISSN: 0747-5632. DOI: 10.1016/j.chb.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0747563217305794.

[66] H. Snyder. "Literature review as a research methodology: An overview and guidelines". In: *Journal of Business Research* 104 (Nov. 2019), pp. 333–339. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2019.07.039. URL: https://www.sciencedirect.com/science/article/pii/S0148296319304564.

[67] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and the PRISMA Group. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement". In: *Annals of Internal Medicine* 151.4 (Aug. 2009). Publisher: American College of Physicians, pp. 264–269. ISSN: 0003-4819. DOI: 10.7326/0003-4819-151-4-200908180-00135. URL: https://www.acpjournals.org/doi/full/10.7326/0003-4819-151-4-200908180-00135 (visited on 01/26/2025).

[68] L.-X. Gao, I. Melero, and F. J. Sese. "Multichannel integration along the customer journey: a systematic review and research agenda". In: *The Service Industries Journal* 40 (Aug. 2019), pp. 1–32. DOI: 10.1080/02642069.2019.1652600.

[69] Clarivate. *2023 Journal Impact Factor, Journal Citation Reports*. 2024. URL: https://jcr-clarivate-com.eaccess.tum.edu/jcr/home.

[70] J. Paul, W. M. Lim, A. O'Cass, A. Hao, and S. Bresciani. "Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR)". In: *International Journal of Consumer Studies* 45 (May 2021). DOI: 10.1111/ijcs.12695.

[71] M. Frasquet, M. Ieva, and C. Ziliani. "Complaint behaviour in multichannel retailing: a cross-stage approach". In: *International Journal of Retail and Distribution Management* 49.12 (2021), pp. 1640–1659. DOI: 10.1108/IJRDM-03-2020-0089. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107515941&doi=10.1108%2fIJRDM-03-2020-0089&partnerID=40&md5=42b55a9d9621ef92f5787921fc56f1bb.

[72] I. Dalla Pozza, A. Brochado, L. Texier, and D. Najar. "Multichannel segmentation in the after-sales stage in the insurance industry". In: *International Journal of Bank Marketing* 36.6 (2018), pp. 1055–1072. DOI: 10.1108/IJBM-11-2016-0174. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047243047&doi=10.1108%2fIJBM-11-2016-0174&partnerID=40&md5=355c229d3c343d396854e4d4abdb3632.

[73] R. Sousa, M. Amorim, E. Rabinovich, and A. C. Sodero. "Customer Use of Virtual Channels in Multichannel Services: Does Type of Activity Matter?" In: *Decision Sciences* 46.3 (2015), pp. 623–657. ISSN: 0011-7315. URL: https://research.ebsco.com/linkprocessor/plink?id=16d1f828-923f-3e25-9774-c9f87e12d1bf.

[74] B. Howcroft, R. Hamilton, and P. Hewer. "Consumer attitude and the usage and adoption of home-based banking in the United Kingdom". In: *International Journal of Bank Marketing* 20.3 (Jan. 2002). Publisher: MCB UP Ltd, pp. 111–121. ISSN: 0265-2323. DOI: 10.1108/02652320210424205. URL: https://doi.org/10.1108/02652320210424205.

[75] L. Rajaobelina and L. Ricard. "Classifying potential users of live chat services and chatbots". en. In: *Journal of Financial Services Marketing* 26.2 (June 2021), pp. 81–94. ISSN: 1479-1846. DOI: 10.1057/s41264-021-00086-0. URL: https://doi.org/10.1057/s41264-021-00086-0.

[76] N. Jo Black, A. Lockett, C. Ennew, H. Winklhofer, and S. McKechnie. "Modelling consumer choice of distribution channels: an illustration from financial services". In: *International Journal of Bank Marketing* 20.4 (Jan. 2002). Publisher: MCB UP Ltd, pp. 161–173. ISSN: 0265-2323. DOI: 10.1108/02652320210432945. URL: https://doi.org/10.1108/02652320210432945.

[77] T.-I. Hu and A. Tracogna. "Multichannel customer journeys and their determinants: Evidence from motor insurance". In: *Journal of Retailing and Consumer Services* 54 (May 2020), p. 102022. ISSN: 0969-6989. DOI: 10.1016/j.jretconser.2019.102022. URL: https://www.sciencedirect.com/science/article/pii/S0969698919309087.

[78] J. Albesa. "Interaction channel choice in a multichannel environment, an empirical study". In: *International Journal of Bank Marketing* 25.7 (2007), pp. 490–506. DOI: 10.1108/02652320710832630. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-35348829255&doi=10.1108%2f02652320710832630&partnerID=40&md5=22b2a00107bbd171fc3cb85addeb0db3.

[79] R. T. Frambach, H. C. Roest, and T. V. Krishnan. "The impact of consumer Internet experience on channel preference and usage intentions across the different stages of the buying process". en. In: *Journal of Interactive Marketing* 21.2 (2007). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dir.20079, pp. 26–41. ISSN: 1520-6653. DOI: 10.1002/dir.20079. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/dir.20079.

[80] S. Sandhu and S. Arora. "Customers' usage behaviour of e-banking services: Interplay of electronic banking and traditional banking." In: *International Journal of Finance & Economics* 27.2 (2022), pp. 2169–2181. ISSN: 1076-9307. URL: https://research.ebsco.com/linkprocessor/plink?id=72275334-a6cb-380e-8329-1a8c6d3b22d4.

[81] Y. Sun, C. Yang, X.-L. Shen, and N. Wang. "When digitalized customers meet digitalized services: A digitalized social cognitive perspective of omnichannel service usage". In: *International Journal of Information Management* 54 (2020). DOI: 10.1016/j.ijinfomgt.2020.102200. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088390456&doi=10.1016%2fj.ijinfomgt.2020.102200&partnerID=40&md5=7d9630935de86e349fe931e94b29c7a6.

[82] K. Jerath, A. Kumar, and S. Netessine. "An information stock model of customer behavior in multichannel customer support services". In: *Manufacturing and Service Operations Management* 17.3 (2015), pp. 368–383. DOI: 10.1287/msom.2015.0523. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84936950413&doi=10.1287%2fmsom.2015.0523&partnerID=40&md5=7e29733ed9334720ea116066d92129e6.

[83] M. Frasquet, M. Ieva, and C. Ziliani. "Understanding complaint channel usage in multichannel retailing". In: *Journal of Retailing and Consumer Services* 47 (2019), pp. 94–103. DOI: 10.1016/j.jretconser.2018.11.007. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057106148&doi=10.1016%2fj.jretconser.2018.11.007&partnerID=40&md5=44be7b2e4150e2d6a781cbfb6f472e31.

[84] X. Ding, R. Verma, and Z. Iqbal. "Self-service technology and online financial service choice". In: *International Journal of Service Industry Management* 18.3 (Jan. 2007). Publisher: Emerald Group Publishing Limited, pp. 246–268. ISSN: 0956-4233. DOI: 10.1108/09564230710751479. URL: https://doi.org/10.1108/09564230710751479.

[85] M. Lipowski and I. Bondos. "The influence of perceived media richness of marketing channels on online channel usage: Intergenerational differences". In: *Baltic Journal of Management* 13.2 (2018), pp. 169–190. DOI: 10.1108/BJM-04-2017-0127. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044780732&doi=10.1108%2fBJM-04-2017-0127&partnerID=40&md5=6d6f842d63f57db1a32558b40f069b5e.

[86] M. Ali, A. Tarhini, L. Brooks, and M. Kamal. "Investigating the situated culture of multi-channel customer management: A case study in egypt". In: *Journal of Global Information Management* 29.3 (2021), pp. 46–74. DOI: 10.4018/JGIM.2021050103. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105103933&doi=10.4018%2fJGIM.2021050103&partnerID=40&md5=9b05cf1b67e5a27b0aeec328901be6ea.

[87] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang. "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts". en. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–21. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581388. URL: https://dl.acm.org/doi/10.1145/3544548.3581388 (visited on 01/25/2025).

[88] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. arXiv:2402.07927. Feb. 2024. URL: http://arxiv.org/abs/2402.07927.

[89] Anthropic. *Review classifier*. en. n.d. URL: https://docs.anthropic.com/en/prompt-library/review-classifier (visited on 01/25/2025).

[90] OpenAI. *OpenAI Platform*. en. n.d. URL: https://platform.openai.com/docs/guides/prompt-generation (visited on 01/26/2025).

[91] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. *Large Language Models are Zero-Shot Reasoners*. arXiv:2205.11916 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2205.11916. URL: http://arxiv.org/abs/2205.11916.

[92] D. Hulbert. *Using Tree-of-Thought Prompting to boost ChatGPT's reasoning*. original-date: 2023-05-22T19:03:27Z. May 2023. URL: https://github.com/dave1010/tree-of-thought-prompting (visited on 02/22/2025).

[93] OpenAI. *OpenAI Platform*. en. n.d. URL: https://platform.openai.com/docs/api-reference/making-requests (visited on 02/22/2025).

[94] H. Aguinis and K. J. Bradley. "Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies". en. In: *Organizational Research Methods* 17.4 (Oct. 2014). Publisher: SAGE Publications Inc, pp. 351–371. ISSN: 1094-4281. DOI: 10.1177/1094428114547952. URL: https://doi.org/10.1177/1094428114547952.

[95] H. Akoglu. "User's guide to correlation coefficients". In: *Turkish Journal of Emergency Medicine* 18.3 (Sept. 2018), pp. 91–93. ISSN: 2452-2473. DOI: 10.1016/j.tjem.2018.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2452247318302164.

[96] J. P. Romano, A. M. Shaikh, and M. Wolf. "Multiple Testing". en. In: *The New Palgrave Dictionary of Economics*. Ed. by Palgrave Macmillan. London: Palgrave Macmillan UK, 2010, pp. 1–5. ISBN: 978-1-349-95121-5. DOI: 10.1057/978-1-349-95121-5_2914-1. URL: https://link.springer.com/10.1057/978-1-349-95121-5_2914-1.

[97] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (Jan. 1995), pp. 289–300. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://academic.oup.com/jrsssb/article/57/1/289/7035855.

[98] R. Delgado and X.-A. Tibau. "Why Cohen's Kappa should be avoided as performance measure in classification". en. In: *PLOS ONE* 14.9 (Sept. 2019). Ed. by Q. Gu, e0222916. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0222916. URL: https://dx.plos.org/10.1371/journal.pone.0222916.

[99] G. Jurman, S. Riccadonna, and C. Furlanello. "A Comparison of MCC and CEN Error Measures in Multi-Class Prediction". en. In: *PLoS ONE* 7.8 (Aug. 2012). Ed. by G. Biondi-Zoccai, e41882. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0041882. URL: https://dx.plos.org/10.1371/journal.pone.0041882.