Services Marketing Quarterly

ISSN: 1533-2969 (Print) 1533-2977 (Online) Journal homepage: www.tandfonline.com/journals/wsmq20

Generative AI Chatbot Prompting for Excellent Customer Service in Tourism

Tommi Kekäläinen, Johanna Heinonen-Kemppi, Juho Pesonen, Erose Sthapit & Brian Garrod

SERVICES MARKETING QUARTERLY
https://doi.org/10.1080/15332969.2025.2610137

Generative AI Chatbot Prompting for Excellent Customer Service in Tourism

Tommi Kekäläinen[a], Johanna Heinonen-Kemppi[a,b], Juho Pesonen[a], Erose Sthapit[c,d] and Brian Garrod[e]

[a]Business School, University of Eastern Finland, Joensuu, Finland; [b]LAB University of Applied Sciences, Lappeenranta, Finland; [c]Department of Marketing, Retail and Tourism, Manchester Metropolitan University, Manchester, UK; [d]School of Hospitality and Tourism Management, Sunway University, Malaysia; [e]School of Management, Swansea University, Bay Campus, Swansea, UK;

ABSTRACT

This article explores the impact of prompting on the performance and service quality of artificial intelligence (AI) customer service chatbots, and the relationship between the prompting process and service quality elements. The data were collected in September 2024 from responses generated by three chatbots, each using a range of prompting techniques. The chatbots responded to real customer inquiries collected by the Visit Helsinki customer service office in summer 2020. The results show that prompting has a significant impact on the quality of online customer service that AI chatbots can provide.

KEYWORDS

Artificial intelligence; chatbots; customer service; prompt engineering; service quality

Introduction

The introduction of digital technologies has transformed how businesses achieve and maintain their competitiveness. One of the best-known and most controversial of such technologies is artificial intelligence (AI), which has found many applications across the business domain. AI chatbots are now widely employed to provide such services (Hsu & Lin, 2023). Studies have shown that AI chatbots can offer many benefits, including the ability to converse naturally and provide personalized recommendations (Dogru et al., 2025). This means that AI chatbots are capable of delivering high levels of service quality, which can positively affect purchase intentions, customer satisfaction, and brand engagement (Blumel et al., 2024).

Chatbots, however, have been found to have a high failure rate, meaning they do not properly understand the customer's request and fail to provide an optimal response. The ability of Chatbots to accurately read customers' emotions and respond accordingly is also vital in customer service contexts (Huang & Rust, 2024). When chatbots fail, this can lead to reduced customer trust and negative word of mouth, which can be reflected in falling sales and fewer repeat purchases (Janssen et al., 2021). The reasons why chatbots fail are not, however, fully understood; nor, therefore, are the best ways to prevent chatbot failure. There is hence a pressing need to evaluate the customer experience of AI and to develop evidence-based remedies (Hsu & Lin, 2023).

Especially novel generative-AI chatbots have the potential to change customer-firm interactions. However, the shift from in-person interactions to automation can be challenging for tourism services (Dwivedi et al., 2024), with a high failure rate a potential barrier. A possible remedy to the tendency for generative-AI chatbots to fail is to fine-tune them using a technique known as prompting. This involves training the chatbot with further background information and example inquiries to optimize its outputs (Henrickson & Meroño-Peñuela, 2025). Robust and reliable studies are, however, needed to explore how prompting might serve to improve the quality of service provided by chatbots and, thereby, the customer experience. AI raises novel issues, such as promises of cost savings and increased

CONTACT Erose Sthapit e.sthapit@mmu.ac.uk Department of Marketing, Retail and Tourism, Manchester Metropolitan University, All Saints, All Saints Building, Manchester M15 6BH; School of Hospitality and Tourism Management, Sunway University, Malaysia.

work efficiency but also concerns about how AI can handle complex interactions in a culturally appropriate way (Wang, 2025).

The purpose of this study is, therefore, to investigate the impact of different prompting techniques on the quality of chatbot responses across various dimensions of customer service in the tourism context. In doing so, the study aims to understand how chatbots can enhance customer service and overall customer satisfaction for tourism businesses. As explained in the following section of the article, the focus will be on how to employ prompting best to enhance the effectiveness of customer-service chatbots used by tourism companies and destinations. Tourism is chosen as the context for this study because it relies heavily on service delivery, which necessitates continuous evaluation and improvement of service quality (Augustyn & Ho, 1998). Tourism is also a particularly complex service-product, where there are ample opportunities for quality gaps to open and for customer service to be required to remedy them (Augustyn, 1998). Finally, tourism is chosen as the context for this study because tourism companies have widely adopted chatbots, which have numerous potential applications in this context (Carvalho & Ivanov, 2024).

Literature review

Customer service in tourism

Tourism is a long-established and highly competitive industry in which service delivery plays a vital role. As tourism products (e.g., hotels, resorts, flights) are often relatively undifferentiated, it can be challenging for a company to make its offering stand out from competitors. One way of doing so is through the services offered to tourists, including the level of customer service (Hudson & Hudson, 2013). At its simplest, customer service is the interaction between the customer and the service provider. This has traditionally been done face-to-face, but increasingly the relationship is being digitalized, including the use of AI-powered chatbots (Carvalho & Ivanov, 2024). Successful customer service is considered a cost-effective way to retain customers, as acquiring new customers is always time-consuming and expensive (Dickson & Huyton, 2008). According to Hennigan (2024), good customer service can increase sales, improve a company's reputation and build long-lasting relationships between customers and businesses. A recent study by Hyken (2024) reports that 88% of customers feel customer service is more important than ever, and 87% feel that good customer service increases trust in a company or brand. Meanwhile, 33% of customers were willing to switch service providers immediately after a poor customer service experience (American Express, 2017).

Service quality

It is widely acknowledged that delivering a high-quality customer experience is an essential ingredient for business success. The term "customer experience" refers to the customer's overall perception and interaction with a brand, company, or organization at all touchpoints throughout their customer journey. Gentile et al. (2007) argue that the traditional drivers of competitiveness, such as price and product quality, are no longer sufficient for most companies: maintaining competitive advantage requires providing excellent customer service. AI is changing digital content creation, but its impact on competitive advantage is still under scrutiny (Guttentag et al., 2025).

While service quality is recognized as playing an important role in determining the overall customer experience, it has proven more difficult to measure service quality than physical product quality. It is possible to examine goods tangibly, based on their various characteristics such as style, color, or feel. Services, however, lack this tangible dimension, making it much more difficult to measure (Grönroos, 1984; Parasuraman et al., 1988). One of the best-known contributions to the understanding of service quality is the SERVQUAL model introduced by Parasuraman et al. (1985). The SERVQUAL model posits that customers assess service quality along five dimensions: reliability, responsiveness, assurance, empathy, and tangibles. It proposes that service quality is determined by the variance between a customer's expectations and their actual perceptions of the service's quality (Wong et al., 1999).

A more recent model for measuring service quality is E-S-QUAL. Developed by Parasuraman et al. (2005), E-S-QUAL is an extension of the traditional SERVQUAL model and is designed to measure the quality of electronic services. E-S-QUAL divides the quality of e-services into four different parts: efficiency (ease and speed of use of the site or service), fulfillment (delivery and availability of the services or products promised by the site to the extent promised), system availability (technical performance of the site or service), and privacy (site security and processing of customer data).

As services increasingly move online and AI is used to deliver them, new requirements for measuring service quality have emerged (Chen et al., 2022). This has led to the development of a model specifically focused on chatbot service quality, called "AI Chatbot Service Quality" (AICSQ). The model recognizes that the characteristics of AI chatbot services differ significantly from human-provided services. For example, an AI chatbot can hold much more information than a human. However, in some areas, such as deep emotional interaction and interpreting emotions, it is far less effective (Chen et al., 2022). Accordingly, Chen et al. (2022) identified seven dimensions to measure service quality provided through AI chatbots: Semantic understanding, Close human–AI collaboration, Humanlike, Continuous improvement, Personalization, Culture adaptation, and Efficiency.

Taken together, these frameworks provide the conceptual backbone for this study. SERVQUAL and related models distinguish between facets of perceived service quality and introduce the concept of expectation–perception gaps, which are central to evaluating AI-based encounters into a broader service journey. Grönroos (1984) distinction between technical and functional quality offers a complementary lens for separating what the chatbot delivers (information accuracy and problem-solving) from how it delivers it (tone, politeness, responsiveness). AICSQ (Chen et al., 2022), meanwhile, highlights chatbot-specific features such as continuous improvement, human–AI collaboration and cultural adaptation. By operationalizing Heinonen and Pesonen's (2022) four elements—effectiveness, responsiveness, politeness and personalization—in the context of generative AI, this study empirically connects the technical concept of prompting to established service-quality constructs.

AI in customer service

Digital customer service

Customer service has traditionally been considered a face-to-face (or at least telephone-based) interaction between a customer and a human customer-service agent or salesperson. Digital transformation, however, has disrupted the status quo, and digital technologies are now being integrated across all areas of business. Customer service is no exception and is increasingly being delivered digitally. Customer preferences and shopping behavior have also shifted significantly toward online services, underscoring the need to invest in various forms of digital customer services (Bacile, 2020; DeLisi & Michaeli, 2021; Lee et al., 2020). The digitalization of customer service has been described as a "win-win-win-win": customers get better, faster service and a better customer experience, and thus stay loyal to the company. This also implies that running the business becomes easier and more profitable (DeLisi & Michaeli, 2021).

Digital customer service is about transforming traditional, analogue customer-to-customer service representative interactions—"moments of truth"—into a digital format: meeting customers where they are (DeLisi & Michaeli, 2021). The platforms where these encounters can take place may include social media (for instance, Facebook, X, or Instagram), the company's own webpage, by e-mail, by text—indeed, by any digital media (Bacile, 2020). Blumel et al. (2024) note that AI-assisted customer service can be divided into three applications. First is conversational analytics, which uses AI to collect and analyze data from various customer service situations to provide feedback to the human customer service agent. Second is conversational coaching, which uses data collected across different customer-service situations to suggest improvements or supplement the human customer-service agent's messages and recommendations. These first two types of service encounters are known as "AI-supported service encounters." Third are chatbots, designed to replace human agents and respond to customer inquiries in a natural, human-like manner. This is known as an "AI-performed service encounter." Even so, chatbots and human customer service agents often share the workload in practice: the chatbot handles simple questions and problems, and the conversation is passed to the human customer service agent if issues arise (Blumel et al., 2024). Hence, while chatbots are becoming increasingly advanced, for the present, they are mainly used as assistants for humans, freeing up time from repetitive, straightforward customer service tasks (De Keyser et al., 2019).

Customer service is about more than simply meeting customers' needs: it is about exceeding their expectations. This helps a company to build an image and a reputation, while also motivating staff and increasing customer satisfaction (Heinonen & Pesonen, 2022). Since today's customers fully expect to receive customer service online (Graef et al., 2021), it is important to understand which elements contribute to online customer-service quality and which detract from it. To this end, Heinonen and Pesonen (2022) analyzed 123 online customer service conversations. They identified four main elements of excellent online customer encounters: effectiveness (informing of delays and providing prompt answers), responsiveness (fulfilling expectations and solving problems), politeness (greeting, thanking, and apologizing), and personalization (providing tailor-made messages to each customer and identifying their personal needs).

Beyond tourism, recent studies in other service sectors show similar patterns in AI-mediated customer encounters. Research on frontline technology use highlights how service robots and AI tools reshape employee roles, service scripts and value co-creation across contexts such as retail banking and healthcare (De Keyser et al., 2019; Graef et al., 2021). Studies on AI chatbots across multiple industries further emphasize that semantic understanding, continuous improvement, and cultural adaptation are key drivers of perceived service quality (Chen et al., 2022). These cross-sector findings suggest that insights from tourism AI chatbots are not isolated but contribute to a broader stream of AI-in-services research.

ChatGPT prompting and programming

The use of AI large language models (LLMs) has developed significantly recently in the field of machine learning (Carvalho & Ivanov, 2024). LLMs can solve a wide range of tasks without being limited to any specific task (Chang et al., 2024; Li et al., 2023). Such models, however, rely on text inputs that may be long and contain spurious information. This can present challenges for language models in tasks that require quick responses or reactions (Li et al., 2023). Customer-service chatbots are a good example of such.

One of the best-known LLMs is ChatGPT. It is a machine-learning software utilizing the Generative Pre-trained Transformer (GPT) developed by OpenAI (Rospigliosi, 2023). The use of ChatGPT is expected to boost productivity and business profitability by automating processes, making them more efficient, and ultimately reducing costs and the need to employ staff. There will also be benefits for users, as services powered by ChatGPT will open up opportunities for efficient, fast, around-the-clock service.

ChatGPT is expected to revolutionize many different business sectors, and the tourism industry is no exception. One of the main uses of ChatGPT in tourism is the personalized advice it can provide. It can, for example, suggest a travel itinerary, including places to visit, eat, and stay—all while taking into account the customer's personal preferences or constraints (Dogra, 2024). ChatGPT can also be particularly useful for various customer service tasks such as handling customer inquiries, assisting in bookings, and managing complaints (Carvalho & Ivanov, 2024).

GPT-4, OpenAI's most recent model, is said to achieve human-level performance across many tasks, including professional applications (OpenAI, 2023). Both the requests and the outputs can be delivered in many world languages (Wu et al., 2023). Despite its usefulness, however, ChatGPT is known to "hallucinate," i.e., to create misleading or entirely false information, which is why businesses should be cautious in simply utilizing ChatGPT for various purposes such as customer service (Carvalho & Ivanov, 2024; Wu et al., 2023).

ChatGPT and other LLMs can, however, be fine-tuned to produce more accurate, relevant, and personalized responses. This requires taking an LLMs and carefully guiding it using "prompt engineering" (Chang et al., 2024). Prompts are textual interactions, questions, statements or other interactions given to LLM, which guide the responses toward a particular outcome (Zamfrescu-Pereira et al., 2023). Optimizing prompts with more context and examples yields better, more accurate responses. Nevertheless, creating effective prompts can be challenging. This is because LLMs are known to be sensitive to any conflicting prompts, causing them to hallucinate (Chang et al., 2024; Zamfrescu-Pereira et al., 2023).

Direct prompting, also known as "zero-shot" prompting, is a simple method that provides specific instructions or questions to the LLM without any background or dataset. The LLM then consults its knowledge base and provides an answer based on the prompt (Zdrok, 2024). When using this prompting technique, the instruction to the chatbot should be as precise and concrete as possible, leaving little room for misinterpretation by the LLM. Henrickson and Meroño-Peñuela (2025) found that adding a short "zero-knowledge" preface to the prompt can improve the accuracy of zero-shot prompting. In practice, this means giving the model a little more context by explicitly describing how it should approach the task (for example: "let us think step by step"). This kind of guidance is beneficial for prompts that involve counting or multi-step reasoning, where language models are prone to making mistakes if they are not encouraged to spell out their reasoning (Kojima et al., 2022).

"Few-shot" prompting builds on zero-shot prompting by conditioning the language model with a small set of illustrative examples (Brown et al., 2020). Because the LLM has already been trained on extensive datasets, these examples help orient the model toward the intended task logic more efficiently (Zdrok, 2024). Few-shot prompting has demonstrated strong performance across many tasks. However, it can be unstable: variations in the choice of examples, their ordering, and the formatting of the prompt can lead to inconsistent outputs (Ma et al., 2023).

Another technique shown to improve language-model reasoning is chain-of-thought (CoT) prompting. CoT instructions encourage the model to break a problem into smaller components and solve them sequentially before producing a final answer, thereby enabling more reliable performance on complex, multi-step tasks. This structured reasoning also makes error diagnosis easier, as it allows identification of the specific step at which a mistaken inference occurred (Wei et al., 2022).

Prompting as a process

Prompting thus involves formulating and presenting a command or instruction to a language model, to improve outputs in a desired way (Zamfrescu-Pereira et al., 2023). Prompting is not, however, simply a technical activity: it can be seen as a process that combines the prompter's own prompt design and the output of a machine (such as a chatbot's response). According to process theory (Van Glabbeek, 2001), a process is the behavior of a system, whether it is a machine, a protocol or, in the present case, a chatbot. Process theory involves two main activities: modeling (representing processes in their own system language) and verification (proving statements about the process, such as whether the system's behavior is as intended). In the prompting process, modeling involves creating prompts, while verification focuses on testing prompting methods and comparing responses to determine their relative quality. By treating prompting as a process, the principles that guide prompt design and implementation can be clarified. In the context of process theory, prompting interactions can be analyzed sequentially. Each prompt leads to a particular output or response, meaning that different prompting methods can be viewed as distinct process "semantics."

Method

This article aimed to examine which of a chosen set of prompting methods are most effective at improving the outcomes of customer-service chatbots in a tourism setting. Based on Heinonen and Pesonen (2022), four key elements of service quality using AI chatbots were assessed: effectiveness, responsiveness, politeness, and personalization. The study used a mixed deductive/inductive approach. This involved coding and categorizing data in pre-assigned categories, with new subcategories emerging as the data were analyzed. In practical terms, we created three versions of the same customer-service chatbot that differed only in how they were instructed (prompted). We then asked each version the same set of real customer questions and compared how well their answers performed on four dimensions of good online service: effectiveness, responsiveness, politeness, and personalization.

Data collection

Based on studies in the general context (Brown et al., 2020; Zdrok, 2024), three techniques were selected for testing in this study: zero-shot prompting, few-shot prompting, and CoT prompting. To guide the style of responses, a role was added at the beginning of all prompts: "You are a Visit Helsinki customer service chatbot. Respond politely and helpfully to customer inquiries in a professional tone using the same language as the question." Customer service chatbots were tested with real questions collected from Visit Helsinki (a Destination Management Organization based in Helsinki, Finland) customer service chat conversations in June and July 2020, using these techniques. Visit Helsinki provided the customer inquiries with organizational permission. All data were fully anonymized prior to analysis, and no personal identifiers were present. As the study used anonymized archival service logs, no separate institutional ethics approval was required under our institution's guidelines.

From a total of 123 chat conversations, 15 different customer questions were purposively selected. We chose inquiries that were relatively long and detailed, and that had elicited what the Visit Helsinki customer service team regarded as high-quality answers. This criterion ensured that each query contained sufficient contextual information to potentially activate all four service-quality dimensions (effectiveness, responsiveness, politeness, and personalization) in the chatbot's responses. However, this purposive sampling strategy prioritizes depth over breadth and may underrepresent shorter, more fragmented or lower-quality customer inquiries.

To ensure a variety of difficulty levels, we selected five questions each from three categories: simple, moderately complex, and complex. Of a total of 41 questions in English and 83 in Finnish, our sample included 4 in English and 11 in Finnish. These conversations were collected during the summer of 2020, during the COVID-19 pandemic, which impacted customer inquiries. Many questions focused on travel restrictions, health regulations, cancellations, and operational changes, resulting in a nature and complexity different from those in typical tourist contexts.

Eleven responses were originally in Finnish and were manually translated into English by a bilingual author. Politeness markers, tone, and stylistic features were preserved to minimize distortion, though we acknowledge that some nuance loss is unavoidable and treat this as a methodological limitation. Another author confirmed that the translations were accurate. Each question was tested separately using three chatbots, each of which was prompted differently. The language model used for the three chatbots was the latest version of ChatGPT (ChatGPT 4o, see OpenAI, 2024). The chatbots will henceforth be referred to as Chatbot A (zero-shot prompting), Chatbot B (few-shot prompting), and Chatbot C (CoT prompting). The questions were asked in their original language, Finnish or English, and the Finnish answers were later translated into English to simplify data analysis. Although care was taken to preserve tone and meaning, this translation step may have influenced how nuances of politeness, personalization, and cultural adaptation were interpreted during coding. However, these factors were considered to the authors' best abilities when confirming the correctness of translations.

All responses were generated *via* the ChatGPT web interface using the default model settings (including the default sampling temperature). For each customer inquiry, we initiated a new conversation and submitted the same role description and task-specific prompt for each prompting condition (zero-shot, few-shot, CoT). The interface does not expose random seeds, and responses are therefore inherently non-deterministic. To mitigate this, we kept all prompt texts constant within each condition and generated all responses within a short time window (September 2024) to reduce potential drift in model behavior. Because random seeds are not accessible and outputs are non-deterministic, each prompt produced one unique response per condition. To limit variability, prompts were kept identical and generated within a short time window. We acknowledge that running multiple generations per prompt would further improve robustness and identify this as future work, but we wanted to simulate real customer encounter circumstances. All prompts that were used to program the chatbots can be found in Appendix 1.

The dataset consisted of 5,746 words. Chatbot A output totaled 2,099 words, Chatbot B 1,578 words, and Chatbot C 2,069 words. The difference in the number of words is not statistically significant (Kruskal-Wallis $p = 0.115$). We acknowledge that the sample is relatively small and purposively selected, and these factors need to be taken into account when interpreting the results.

Data analysis

Quantitative content analysis (QCA) was chosen as the research method. Despite its name, it is a qualitative research method useful for identifying trends and patterns in large textual datasets (Brazzoli, 2023). QCA is a systematic and objective procedure for describing communication, including segmenting data into units and categories and later creating summaries of these categories (Rourke & Anderson, 2004). In QCA, coding involves categorizing the data, i.e., using codes to summarize and label the data (which could be a word or phrase), into specific notes or memos. This process allows further analysis or visualization of the data (Linneberg & Korsgaard, 2019). The aim was to identify differences in responses across chatbots and determine which chatbot performed best across all four customer service dimensions.

The performance of chatbots was assessed based on the key elements of excellent online customer encounters identified by Heinonen and Pesonen (2022): effectiveness, responsiveness, politeness, and personalization. In their original study, the researchers applied these elements to human-to-human online customer service interactions. In this study, we adapt the same four elements to the context of AI-powered customer service. To capture chatbot-specific features, we divided each element into subcategories (see Figure 1). Some subcategories stem from Heinonen and Pesonen (2022), while others were developed inductively from the data.

This study refines the concept of effectiveness. While response speed is a key indicator in human service encounters, it is a fundamental feature of AI chatbots. Thus, we defined effectiveness as “Proactive communication” and “Informing of delays,” alongside accuracy, relevance, and completeness of information, all of which are crucial for effective service delivery. Although “Informing of delays” was rarely observed, it was retained for theoretical consistency with earlier frameworks.

The responses from the three chatbots were uploaded to Atlas.ti, a software package commonly used for qualitative data analysis (Paulus et al., 2019). Data collected from chatbot responses were coded by carefully reading the data phrase by phrase, assigning each line to a category and then a specific subcategory. Phrases or lines that were ‘empty’ in content or did not fit into any category were ignored. Finally, the quantified data were collated into a table, which enabled evaluation of the occurrence of different codes and comparison of the performance of the three chatbots. A single researcher first performed coding. Then, an independent coder performed another round of coding. We followed the intercoder reliability process guidelines outlined by MacPhail et al. (2016). Intercoder reliability was assessed retrospectively on 10 double-coded transcripts (22% of the sample) using Cohen’s Kappa. Overall agreement was substantial ($\kappa = 0.78$, 95% CI: 0.72–0.85). Reliability varied by dimension: excellent for Politeness ($\kappa = 0.95$) and Personalization ($\kappa = 0.81$), but poor for Responsiveness ($\kappa = 0.00$) and Effectiveness ($\kappa = 0.17$). The notes between coders were compared, and we identified the difference to be due to systematic coding differences. The differences were systematically compared, and changes were made to the coding manual according to the observations, especially regarding Responsiveness and Effectiveness. The second coder then re-coded the complete data set according to the revised version of the coding manual. The main difference is in the Responsiveness dimension, where the new analysis identified significantly more instances, particularly for Chatbot A (39 vs. first coding 24). Additionally, the new coding yielded a higher total count (235 vs. 217) due to a more frequent application of problem-solving codes, although the original finding that Chatbot C dominates in Personalization and Politeness remained consistent. The final results are reported in Table 2, summarizing the number of discussions where different dimensions were observed.

Finally, quantitative comparative analysis was conducted to compare the differences between chatbots regarding their use of words and how different dimensions were represented in the data set. Statistically significant differences allow more robust conclusions from the data.

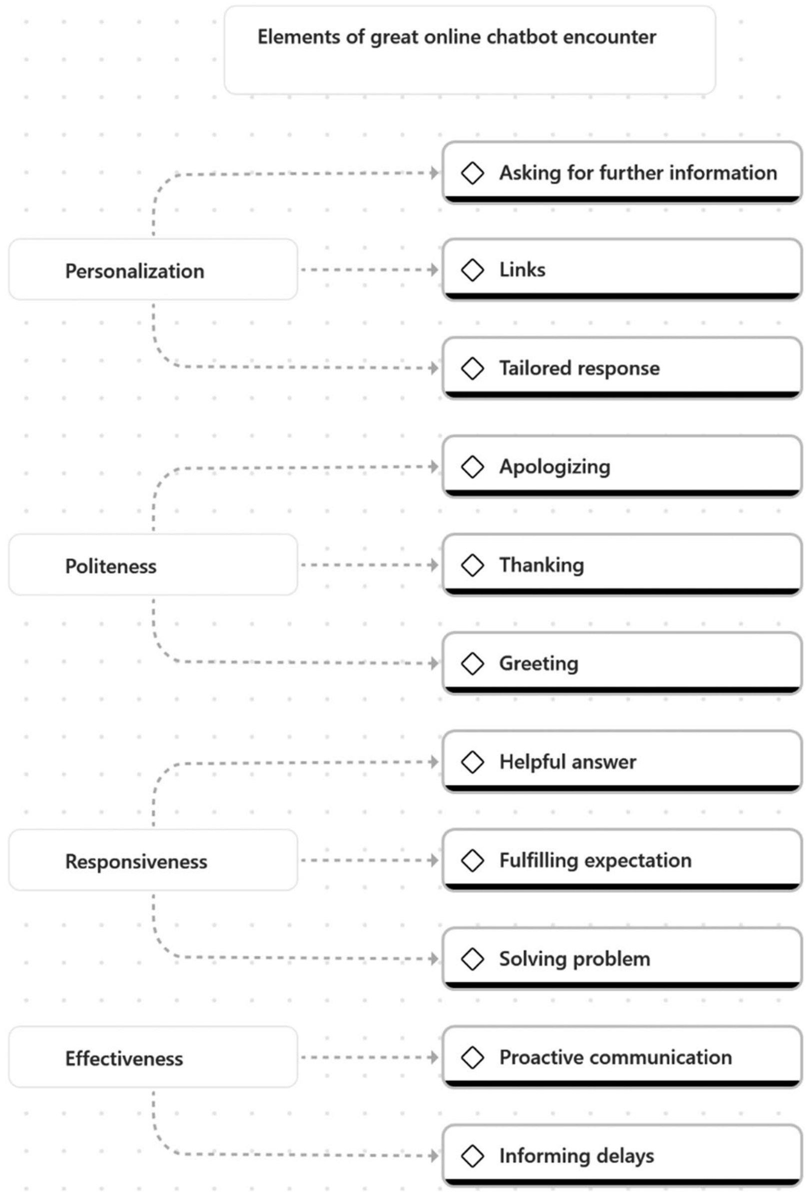

Figure 1. Elements of excellent online chatbot encounter adapted from Heinonen and Pesonen (2022).

Table 1. Summary of common prompting methods based on earlier studies.

Prompting method	Findings in earlier research	Reference
Zero-shot prompting	Involves a specific question or instruction to the language model without background knowledge. Efficient for straightforward tasks.	Dang et al. (2022) Zdrok (2024)
Few-shot prompting	Conditioning the language model with a small number of examples, helping to approach the task from the right angle. Effective but suffers from instability due to different variations of the examples.	Brown et al. (2020) Ma et al. (2023) Zdrok (2024)
Chain-of-thought prompting	Breaking down the task into several steps. Found to be effective in more demanding problems. Allows tracing the inference process and easier troubleshooting.	Wei et al. (2022)
Instruction prompting	Involves clear, explicit instructions for the language model, aiming to reduce misunderstandings. The strength of this method is the expressive power of clear instructions compared to examples.	Efrat and Levy (2020) Giray (2023)
Retrieval-Augmented prompting	Allows the language model to access knowledge from external sources such as company's own databases. Provides reliable information for chatbot, thus reliable outputs.	Gao et al. (2023) Lohrbeer (2023)
Role prompting	Adding a persona to the prompt. Effective in maintaining specific tone and guiding the style of the outputs.	Kong et al. (2023) Wang et al. (2023) Zheng et al. (2023)

Results

This section of the article presents the results on the performance of the three chatbots across the four dimensions of good customer service. Fifteen questions were tested with each of the three chatbots, yielding 45 responses. The questions can be roughly divided into four main groups: practical information and local regulations (4 questions), activities and attractions (4 questions), local venue recommendations (4 questions), and transportation and accommodation (4 questions).

Chatbot performance

Overall, the chain-of-thought prompted chatbot produced the wealthiest and most personalized answers; the zero-shot chatbot gave shorter, more concise answers but was less polite; and the few-shot chatbot generally fell between these two. The following subsections provide a more detailed breakdown of these differences.

After the coding process, the frequency of each code in responses from the different chatbots was quantified. From the number of codes, the chatbot's performance across different dimensions of customer service can be inferred. The detailed breakdown of the performance of each chatbot is presented in Table 2. In the table, the number of discussions where the dimension quotations covered by a code is reported except for "Links" where the total number of links in the discussions is presented.

"Personalization" included "Asking for further information," "Links (provided)," and "Tailored response" to the customer. In practice, this meant that the customer service agent, in this case the chatbot, would make a reasonable effort to find information tailored to the customer's specific needs. "Politeness" focused on basic polite expressions provided by the chatbot, including Greeting, Thanking, and Apologizing. "Responsiveness" included "Helpful answer," "Fulfilling expectation," and "Solving problem." The purpose of this category was to identify whether the response was helpful or whether the customer received an answer or solution to their problem or question. Heinonen and Pesonen (2022) defined "effectiveness"" mainly in terms of speed of responses and keeping the customer informed throughout the conversation. With chatbots, the responses were generally instant, so this category was adjusted to focus on "Informing of delays" and "Proactive communication." As such, this category was used to assess how well the chatbot notified users of potential response delays and proactively kept customers updated.

Classification of coded data

Table 1 shows the performance of different chatbots in different dimensions of customer service. For each chatbot and service-quality dimension, we first summed the number of quotations coded under all relevant subcategories. These raw frequencies form the numeric scores reported in Tables 2–4.

Table 2. Number of discussions where each dimension is identified.

	Chatbot A	Chatbot B	Chatbot C	Totals
Personalization	11	7	27	45
Asking for further information	5	3	15	23
Links	1	2	3	6
Tailored response	5	2	9	16
Politeness	12	24	30	66
Apologizing	0	0	0	0
Thanking	3	10	15	28
Greeting	9	14	15	38
Responsiveness	39	35	35	109
Helpful answer	10	5	5	20
Fulfilling expectation	14	15	15	44
Solving problem	15	15	15	46
Effectiveness	2	4	9	14
Informing delays	0	1	1	2
Proactive communication	2	3	8	13
Totals	64	70	101	235

The chatbots' performance scores ranged from 0 to 39. Five categories were developed to reflect different levels of performance. Specifically, the 0–39 range was divided into five equal-width bands to obtain interpretable qualitative labels: 0–8 as Poor, 9–16 as Moderate, 17–24 as Good, 25–32 as Very Good, and 33–39 as Excellent (Table 3). These labels are intended as heuristic descriptors to aid interpretation rather than statistically derived cutoff points. This made it easier to describe the strengths and weaknesses of each chatbot: for example, which was best at giving personalized responses or which was most polite (Table 3).

Findings

Summary of chatbot performance in customer-service dimensions

By combining the results from Table 2 and the code-range descriptions from Table 3, the chatbot responses were presented in an easy-to-read format (Table 4). Chatbot C received higher scores than Chatbot A and Chatbot B on most coded dimensions. The analysis of Chatbots A and B yielded almost identical numbers of discussions (64 for A and 70 for B), with some differences across the customer service dimensions. At the same time, Chatbot C scored significantly higher with 107 different codes in discussions.

Personalization

Chatbot C was the best performer in the "Personalization" dimension, receiving a rating of "Very Good" (27 coded discussions), reflecting its ability to produce well-tailored answers to questions. Both Chatbot A and B had a "Moderate" performance in "Personalization." However, the differences between these two chatbots were relatively small, with the most noticeable difference being in the "Tailored response" sub-category, where Chatbot A scored five coded discussions and Chatbot B scored only two. Chatbot B was also more likely to present links than Chatbot A.

Chatbot C excelled especially in "Asking for further information" at the end of each answer to find out whether there was any other way it could help:

> *… I can look up more specific details for you. Let me know if you need any further assistance!* (Chatbot C response to Question 1).

Chatbot C also searched and provided the most direct links to support its suggestions, therefore being the best performer in the subcategory "Links." It used links a bit more often than others with higher frequency:

> *… For transportation, you can consider the **HSL 3-day ticket [hyperlink]**, which provides unlimited travel on public transport within the Helsinki area. Here are a couple of links where you can read more …* (Chatbot C response to Question 5).

Table 3. Descriptions of code ranges.

Code range	Description
0–8	Poor
9–16	Moderate
17–24	Good
25–32	Very Good
33 or more	Excellent

Table 4. Summarized chatbot performance in customer service dimensions.

Customer service dimension	Chatbot A	Chatbot B	Chatbot C
Personalization	Moderate	Moderate	Very Good
Politeness	Moderate	Good	Very Good
Responsiveness	Excellent	Excellent	Excellent
Effectiveness	Poor	Poor	Moderate

This supports the perception that the customer-service chatbot is making an effort to find a good solution and gather more information (Heinonen & Pesonen, 2022). Chatbot A and Chatbot B both provided links that were poorly supported in their answers, and both received a “Poor” rating.

The subcategory “Tailored response” was intended to assess whether the chatbot’s response was focused on the specific question and not overly generic. For instance, Question 6 was about finding a place for a birthday party, where the chatbot began its answer by:

> *Thank you for your message, and congratulations on your upcoming birthday!* (Chatbot A response to Question 6).

This demonstrated situational awareness and the ability to tailor the answer to the question. Another good example of a tailored response was Chatbot C’s answer to Question 4, when asked about a good place to go with children in variable weather:

> … I understand you are looking for a nice spot to spend time with children in unstable weather. Here are a few suggestions in Helsinki: 1. SEA LIFE Helsinki—This aquarium is perfect for children of all ages, and indoors, you can explore the wonders of the underwater world. 2. Leikkiluola—An indoor play area located in the Itäkeskus shopping mall, offering plenty of space for playing, climbing, and burning off energy. It is great for children aged 1–10.

Instead of simply suggesting an indoor venue, the chatbot found places that were especially suitable for children.

Politeness

All chatbots performed reasonably well in “Politeness,” but there were some apparent differences in the results between the different chatbots. Again, Chatbot C received the highest possible rating of “Excellent” in this dimension, with a total of 30 coded discussions, along with Chatbot B, which received 24. Chatbot A, meanwhile, received only 12 coded discussions in this dimension.

None of the chatbots apologized in their responses for any questions, likely because each conversation consisted of only one question and one answer. A situation in which an apology would have been appropriate did not therefore arise, resulting in a total of 0 coded discussions for “Apologizing.” Chatbot C excelled in both “Thanking” and “Greeting,” and scored 15 coded discussions, thanking the user in each response. Chatbot B similarly greeted in each response but one, receiving 14 coded discussions in total. However, it only thanked 10 times, sometimes failing to thank the user for the question.

> *Hello, and thank you for your question!* (Chatbot C response to Question 1).

> *Hello! For souvenir shopping in Helsinki, you have several great options …* (Chatbot B response to Question 2).

Chatbot A raised only three coded discussions for “Thanking” and nine for “Greeting,” making it the weakest in this category.

Responsiveness

A highly responsive chatbot will provide accurate responses to user queries, solving the customers’ problems and fulfilling their expectations (Heinonen & Pesonen, 2022). The results for the “Responsiveness” category were the most balanced among the three chatbots, with Chatbot A receiving altogether 39 coded discussions, Chatbot B receiving 35, and Chatbot C receiving 35. This resulted in a rating of “Excellent” for A, B, and C.

The subcategory of “Helpful answer” was to assess whether there was any extra information (something that was not directly asked) or tips in the response that could be helpful, such as local rules, guidelines, or regulations. Here, Chatbot A performed best, often giving tips such as:

> *… Keep in mind that the nature on the island is quite sensitive, so littering and damaging the environment are strictly prohibited. It is also a good idea to check any specific rules or opening hours on the City of Helsinki’s website before your visit.* (Chatbot A response to Question 7)

Chatbot A received 10 coded discussions in this subcategory. However, Chatbot B (five coded discussions) and Chatbot C (five coded discussions) were not that far behind, both often giving similar tips in their responses to Chatbot A.

'Fulfilling expectation' also received a very similar number of coded discussions across all three Chatbots. In contrast to the study by Heinonen and Pesonen (2022), in which the discussion outcome was used to determine whether expectations were met, here the researchers had to make their own assessment. In practice, the goal was to evaluate whether the chatbot's response provided sufficient information. Here, Chatbot A received 14 coded questions, while Chatbot B received 15, and Chatbot C received 15.

The key difference between 'Fulfilling expectation' and 'Solving problem' was that the latter required the question to be answered directly and clearly, meaning that the exact phrase could sometimes be coded using both codes. For example, when asked *"Do you know if in Finland is legal to make barbecues in parks?" (Question 1),* the response from Chatbot B was:

> *…Yes, in Finland, it is generally legal to make barbecues in designated public park areas, as long as there is no specific ban due to fire risk or local regulations.*

This answer provided a direct solution to the problem while also fulfilling the customer's probable expectations.

Effectiveness

In terms of "Effectiveness," Chatbot C performed the best with 9 coded discussions, receiving a rating "Moderate," followed by Chatbot B, which had a total of four coded discussions with a rating "Poor," and Chatbot A, with two coded discussions, resulting in a "Poor" rating. The chatbots generally did not inform users about delays in responses, with Chatbots B and C each doing so only once, while Chatbot A did not do so at all. This was to be expected, as chatbot responses tend to be immediate, so there is no necessity to inform the user of a delay.

When analyzing the chatbot responses, the second identified sub-category was "Proactive communication," which refers to how the chatbot kept its responses up to date. In practice, this was reflected in the following example:

> *Is Seurasaari open today?* (Question 3)
>
> … *I'll check the opening hours for Seurasaari right away.* (Chatbot C response)

In this subcategory, Chatbot C performed best with 8 coded discussions, while Chatbot A received three and Chatbot B received two.

Statistical comparisons

Quantitative examination of chatbots show that Chatbot B (few-shot) generated the most concise responses with a mean of 105.2 words (median = 96.0), ranging from 55 to 216 words, and demonstrated the most consistent performance with a standard deviation of 41.9 words. In contrast, Chatbot A (zero-shot) and Chatbot C (chain-of-thought) produced notably longer responses with nearly identical means of 139.9 and 137.9 words, respectively, and both sharing a median of 129.0 words. However, Chatbot A exhibited substantially greater variability ($SD = 65.9$) with responses ranging from as brief as 49 words to as extensive as 308 words, while Chatbot C showed more moderate variability ($SD = 48.5$) with a range of 69 to 243 words. Overall, the few-shot approach produced responses approximately 25–30% shorter on average compared to the zero-shot and chain-of-thought approaches, while also maintaining greater consistency in response length. However, the differences were not statistically significant.

We also compared the differences in the number of codes in different dimensions between chatbots. The frequency-based analysis of 10 service quality code categories (omitting "Apologizing") across three chatbot types revealed that four categories demonstrated statistically significant differences. "Thanking" ($p < 0.001$), "Asking for further info" ($p < 0.001$), "Greeting" ($p = 0.001$), and "Proactive communications" ($p = 0.038$) frequencies differed significantly. This indicates that prompting strategy affects conversational richness and engagement frequency rather than fundamental service competence.

The Kruskal-Wallis test revealed significant differences in overall code frequency between chatbot types ($H = 25.35$, $p < 0.001$), demonstrating that prompting strategy fundamentally affects response comprehensiveness measured by total code occurrences. Chatbot C prompting produced the most code-rich responses with an average of 6.80 codes per response (median = 7.0, $SD = 1.01$, total = 109 codes across 15 responses), representing approximately 59% more codes than Chatbot A (mean = 4.27, median = 4.0, $SD = 1.03$, total = 64) and 34% more than Chatbot B (mean = 5.07, median = 5.0, $SD = 1.33$, total = 76).

Discussion

As encounters between service providers and customers increasingly take place online and customer service is increasingly performed by AI chatbots (Berg et al., 2022), it is vitally important that companies know how to achieve and maintain excellent service quality. This article has sought to develop such knowledge by evaluating the prompting process and the performance of three differently prompted chatbots across four previously identified dimensions of excellent online customer service. A set of 15 questions was tested with each chatbot to assess how well they perform in real-life situations across each dimension and to determine their relative effectiveness.

Prompting process

Prompting is a multi-stage process, and meeting the characteristics of excellent customer service requires examining various service quality theories. Table 5 presents a mapping of the prompting process stages to service-quality dimensions and their linkage to process theory.

The first phase of the prompting process involves branching, which, in process theory, refers to the state of the process at which multiple choices are possible (Van Glabbeek, 2001). There are various methods of prompting (also known as initialization), three of which have been applied in this study. The results suggest that no single prompting method is superior across all service-quality elements. The appropriate prompting method will depend on the tasks the Chatbot is being put to and the particular dimension or dimensions of excellent service quality the company wishes to prioritize.

The next phase of the prompting process is modeling, which involves creating prompts that are believed to produce reasonable responses. Verification is then required, which involves testing different prompting methods to determine which leads to the best-quality responses. As such, they are considered central to the process theory (Van Glabbeek, 2001). This part of the process relates to functional quality, as Grönroos (1984) defines it, the service-quality dimension that corresponds to how service quality is perceived.

The prompting process also involves troubleshooting and correcting errors to ensure the chatbot's outputs remain consistently high quality. This aligns with the process theory element known as "failure tracing," which involves observing the process to understand why the system failed under certain conditions (Van Glabbeek, 2001). Regarding service-quality theories, this part of the process relates to the AICSQ theory dimension "continuous improvement" (Chen et al., 2022), which enhances system stability and improves its capabilities.

Table 5. Prompting process stages linked to process theory and service-quality dimensions.

Process theory concept	Service quality element	Prompting process stage	Connection to process theory
Branching	Reliability	Initialization	Ensuring each step of the process is consistent and accurate contributes to reliable service. This stage of the prompting process involves choosing appropriate prompting methods and designing prompt.
Modeling Verification	Responsiveness Efficiency Functional quality	Creating prompts Testing	Modeling prompts by creating an effective prompt in system language and verification by testing its functionality, contributing to a system's responsiveness.
Failure tracing	Continuous improvement	Troubleshooting Fixing errors	Testing the prompt and troubleshooting errors to ensure the system functions reliably. Continuously improving the system with updates.
Action Relations	Tangibles Technical quality	Output presentation	Actions taken in the process affect the tangible outputs, how one action leads to a specific change in the output. Effects the appearance and structure of the responses.

The final phase of the prompting process concerns the prompt's output and its presentation. It is the tangible, most visible part of the process, shaped by the various actions and transitions throughout, and is also a key dimension of process theory. This relates to how processes evolve: performing a certain action transitions a process to another, thereby affecting the result (Van Glabbeek, 2001). The chatbot prompting process thus embodies the system's ability to transition from a prompt to concrete, understandable output. This part of the process can be viewed as its technical quality, as defined by the SERVQUAL model, i.e., what the customer receives when interacting with the service (Grönroos, 1984).

Prompting methods

In the foregoing analysis, Chatbot C, which used the CoT prompting method, achieved the highest scores across almost every dimension of excellent online customer service in our dataset. In this exploratory, small-sample study, this pattern suggests that a chatbot instructed with a CoT-style prompt can deliver relatively higher-quality customer service for the kinds of complex, information-rich inquiries analyzed here. Statistically examining the coding result differences suggests a hierarchical model of service quality enhancement measured through code frequency: Zero-shot provides baseline problem-solving functionality with moderate code usage and moderate variability (range 2–6 codes per response), Few-shot adds consistency in basic social behaviors while maintaining efficiency (range 4–9 codes), and Chain-of-thought maximizes comprehensiveness by incorporating explicit reasoning steps that encourage fuller engagement, more frequent use of politeness codes, more transparent communication, and more thorough follow-up behaviors (range 6–10 codes). Thus Chain-of-thought ultimately creates richer and more complete service interactions through higher code frequency despite requiring more computational resources. However, given the limited and non-representative sample, this should be treated as a context-specific indication rather than a general claim that CoT prompting is universally superior.

Chatbot C delivered a very high level of personalization, responding in a specific and tailored manner to the questions and demonstrating context awareness. Chatbot B (which used few-shot prompting) performed reasonably well, ranking first in responsiveness (where all chatbots performed well), but lagged well behind Chatbot C overall. Chatbot A (which used zero-shot prompting), meanwhile, produced the worst results overall, being responsive but struggling to deliver politeness.

The results show that prompting methods have a substantial impact on the behavior of customer service chatbots. Although AI language models cannot be fully controlled, it is possible to guide their tone and emphasis on different customer service dimensions by providing them with background knowledge through prompting. A summary of the strengths and weaknesses of different prompting methods is presented in Table 6.

Overall, the CoT method performed best, but the zero-shot method also produced concise, relevant responses. The performance of the few-shot prompting method, meanwhile, depended mainly on the quality of the examples given and was therefore somewhat difficult to evaluate. However, in objective terms, it did perform reasonably well in this study.

From a practical perspective, the three prompting strategies can be seen as tools for different parts of the customer-service workflow rather than competing universal solutions. Zero-shot prompting is

Table 6. Prompting methods' strengths and weaknesses.

Prompting method	Advantages	Weaknesses
Zero-shot Prompting	- Ease of use, no prior examples or guiding needed - Responsive enough, useful for general and short queries	- Lacking politeness and personalization - Limited in handling multiple-part questions
Few-shot Prompting	- Providing examples increases chatbot politeness - Excellent responsiveness	- Requires more set-up time with long prompts - Response quality can depend on the examples provided
Chain-of-Thought Prompting	- Step-by-step reasoning leads to highly personalized, rich and long responses - Logical and coherent responses	- May result in too much information in responses to simple queries - Set-up and breaking down the reasoning might be confusing

well-suited to high-volume, low-complexity interactions, such as answering frequently asked questions or confirming opening hours, where speed and brevity are more important than depth or emotional nuance. Few-shot prompting becomes useful when tourism organizations wish to impose a recognizable "house style" or policy logic on the chatbot, for instance, by providing examples of how to handle complaints, respond to sensitive topics (e.g., cancellations, health restrictions), or prioritize certain products. CoT prompting is most appropriate when queries require the chatbot to combine several pieces of information (e.g., planning an itinerary within budget and time constraints, explaining layered regulations, or balancing different customer preferences) and where more elaborate justification is likely to be appreciated by the customer.

An analysis of weaknesses in prompting methods provides key insights. The limited politeness in zero-shot prompting arises from the model defaulting to an informational style without specific tone instructions. This leads to efficient but socially blunt responses. Variability in the few-shot method depends on the quality of examples provided. Strong examples lead to contextually rich responses, while poor examples result in generic outputs. Although CoT prompting produces the best results, its step-by-step reasoning can result in excessively long responses for simple queries. Therefore, prompt engineers should choose methods based on their strengths while being aware of potential weaknesses.

In operational terms, this means that tourism organizations should avoid a single, "one-size-fits-all" prompting strategy. Instead, prompting needs to be matched to the typical service scenarios in which the chatbot operates. For example, a destination management organization may deploy zero-shot prompting for short, transactional queries handled directly on the homepage, while routing complex travel-planning questions to a CoT-prompted instance that is explicitly instructed to reason step by step and verify information before answering. Few-shot prompts can be reserved for situations where consistency with brand guidelines is critical, such as responding to complaints or explaining refund policies. This layered use of prompting strategies can maximize service quality while keeping computational and design efforts under control.

Ethical considerations in AI-powered customer service are crucial. Customers often experience discomfort or distrust when interacting with non-human agents, a phenomenon described as ethical anxiety (Wang & Zhang, 2025). Human–AI interaction research similarly points to the importance of clarity about system capabilities, data usage and the division of labor between humans and AI in service encounters (Graef et al., 2021). In our context, this implies that prompting strategies should be designed alongside transparent data governance and transparency practices: customers should understand when they are interacting with an AI, how their data is used, and under what conditions conversations are logged, reviewed, or escalated to humans. Even excellent technical responses may not guarantee a positive experience if customers perceive the AI as opaque, unaccountable or unfair.

Additionally, the connection between prompting methods and chatbot service failure is significant. Misunderstandings and "hallucinations" are common challenges. The chain-of-thought method can reduce these issues and improve the handling of complex queries. In contrast, zero-shot prompting can oversimplify, increasing the risk of miscommunication. The effectiveness of the few-shot method relies on high-quality examples. Overall, prompt engineering serves not only to enhance service quality but also to minimize the risk of critical chatbot failures.

For service marketing, the main implication is that prompting can be treated much like scripting and training human frontline staff: by carefully designing prompts, organizations can directly influence how trustworthy, helpful, and personalized AI-mediated encounters feel to customers.

Interpreting the findings through service quality and service encounter theory

The comparative performance of the three prompting methods can be examined through established service-quality and service-encounter frameworks. Building on SERVQUAL's foundational contribution to understanding service quality (Parasuraman et al., 1985), the four dimensions used in this study—effectiveness, responsiveness, politeness and personalization (Heinonen & Pesonen, 2022)—map closely onto SERVQUAL's original dimensions: effectiveness and responsiveness reflect reliability and responsiveness, politeness reflects assurance and empathy, and personalization is linked to empathy through recognition of individual needs. The superior performance of CoT prompting in politeness and personalization therefore indicates that prompt design can be used as a lever to enhance the “soft” relational dimensions of service quality (Parasuraman et al., 1985), not only its technical accuracy.

These findings are particularly relevant when viewed through Grönroos (1984) distinction between technical quality (what the service delivers) and functional quality (how it is delivered). CoT prompting appears to enhance both dimensions: the step-by-step reasoning improves technical quality by making the decision process more transparent and accurate, while the structured approach simultaneously improves functional quality through more polite and personalized interactions. This dual enhancement is critical in the context of AI-delivered services, where the intangible nature of service quality makes measurement inherently more difficult than for physical products (Grönroos, 1984; Parasuraman et al., 1988).

The results also align with Chen et al. (2022) AICSQ model, which recognizes that AI chatbot service quality differs significantly from human-provided services. Of AICSQ’s seven dimensions the present findings most directly address semantic understanding, human-like characteristics and personalization. CoT prompting’s superior performance in politeness and personalization suggests that this prompting approach better approximates the human-like qualities that customers value in AI interactions (Chen et al., 2022). The structured reasoning process embedded in CoT prompts may help bridge the gap between AI’s superior information-processing capabilities and its traditionally weaker performance in emotional interaction and empathy interpretation (Chen et al., 2022).

Furthermore, these findings illuminate the expectation-perception gap that is central to service quality evaluation (Parasuraman et al., 1985; Wong et al., 1999). As AI increasingly delivers services traditionally provided by humans (Guttentag et al., 2025), customers bring expectations shaped by both human and digital service encounters. CoT prompting may help narrow this gap by enabling chatbots to meet the higher-order expectations associated with personalized, empathetic service delivery. This capability is particularly important as companies seek to maintain competitive advantage through excellent customer service in an era where AI is transforming digital content creation and service delivery.

Conclusions

Theoretical and methodological contribution

The main theoretical contribution of this article extends service-quality theory by demonstrating that prompt structure systematically influences AI chatbot performance across established service-quality dimensions. Specifically, the superior performance of CoT prompting in politeness and personalization—dimensions traditionally associated with human relational skills—challenges the assumption that AI systems are inherently limited in “soft” service attributes (Chen et al., 2022). By mapping Heinonen and Pesonen’s (2022) four elements onto SERVQUAL’s dimensions and empirically testing three prompting methods, this study reveals that prompt engineering functions as a design lever for both Grönroos (1984) technical quality (what is delivered) and functional quality (how it is delivered). This finding theoretically advances AICSQ (Chen et al., 2022) by showing that dimensions like “human-like characteristics” and “personalization” are not fixed chatbot limitations but malleable through prompt design.

The study also contributes to process theory (Van Glabbeek, 2001) by reconceptualising prompting as a multi-stage process analogous to service-encounter scripting. The empirical differences between zero-shot (task-focused), few-shot (example-based) and CoT (reasoning-exposed) methods map onto distinct process configurations: branching determines initial strategic direction, modeling embeds behavioral templates, verification ensures output alignment, and failure tracing enables continuous improvement. This framing bridges abstract process theory with concrete AI service design, demonstrating how “backstage” prompt engineering shapes “frontstage” customer interactions. This mechanism has been previously undertheorised in AI-mediated service encounters. The systematic performance variation across prompting methods confirms that the expectation-perception gap (Parasuraman et al., 1985) can be narrowed through intentional prompt design, offering a novel theoretical pathway for enhancing AI service quality.

Methodologically, this article contributes by operationalizing service-quality evaluation through quantitative content analysis of AI responses to authentic customer questions. This approach moves beyond hypothetical scenarios or user satisfaction surveys, enabling direct assessment of chatbot output quality against established service dimensions. The use of real customer queries ensures ecological validity while the comparative design isolates the causal effect of prompting method, thus advancing methodological rigor in AI service research.

Managerial implications

This study offers managers important insights into how to enhance AI-driven customer service chatbots through effective prompting strategies. It demonstrates that specific prompting methods influence chatbot effectiveness, responsiveness, personalization, and politeness, enabling organizations to tailor chatbots to meet their needs. Rather than viewing prompting as a technical task, it should be regarded as a strategic design choice that enhances customer experience. For instance, if a company wants to convey a friendly image, its chatbot can be programmed to adopt a friendly tone; if a formal image is desired, the chatbot can be prompted accordingly. The study also reveals that the CoT prompting method generates longer, more detailed responses suitable for complex inquiries, while zero-shot prompting yields shorter, straightforward answers for simpler questions. The success of chatbots depends on the surrounding organizational ecosystem, requiring robust support, ongoing training, and skill development. The three prompting strategies (as outlined in Table 7) serve different purposes in customer service workflows. Organizations should categorize customer inquiries by complexity; for example, zero-shot prompting works best for high-volume factual questions, while CoT prompting is ideal for complex, multi-step inquiries. In practice, organizations can operationalize these strategies through routing rules within chatbot platforms. Responsibilities for designing prompts, reviewing conversation logs, and refining techniques should be clearly assigned to ensure effective implementation. While CoT prompting offers richer responses, it incurs longer response times and higher computational costs, presenting a quality-latency tradeoff. CoT prompting can also enhance service design and employee training by aligning chatbot responses with existing decision-making processes. This approach supports organizational learning and makes tacit service principles explicit. However, organizations must also manage risks, as no prompting method is without drawbacks.

More generally, implementing chatbots can improve a company's efficiency by automating labor-intensive tasks. Cacic (2023) explains that by choosing the appropriate prompting method and fine-tuning it to the company's specific needs, it is possible to handle complex cases automatically, saving time for a human customer service agent. When a chatbot has not been fine-tuned, the human customer service agent may find themselves handed cases that have become complicated due to a misunderstanding. This can be frustrating for both the agent and the customer. Automation also speeds up response times and can ultimately improve customer satisfaction.

Table 7. Managerial decision framework for selecting and implementing prompting strategies in AI-based tourism customer service.

Decision area	Managerial action	Practical guidance
1. Define the service objective	Match the prompting method to the interaction's primary goal.	**Zero-shot:** maximize speed and efficiency for routine factual inquiries. **Few-shot:** ensure consistent tone, style, and policy adherence in moderately complex situations. **Chain-of-thought:** provide detailed, step-by-step guidance for complex, multi-layered, or sensitive inquiries.
2. Specify escalation rules	Establish clear criteria for when conversations should be transferred to human agents.	Embed escalation triggers directly in system prompts (e.g., safety concerns, complaints, regulatory questions, customer distress). Ensure the chatbot explicitly acknowledges uncertainty and directs users to human support when needed.
3. Monitor service quality	Track chatbot performance systematically and revise prompts when issues arise.	Evaluate both technical metrics (e.g., response speed) and service-quality indicators aligned with this study: personalization, politeness, responsiveness, and perceived effectiveness. Conduct periodic expert reviews and update prompts accordingly.

Limitations and suggestions for future research

This study has several limitations that may affect the results. The small sample size could compromise reliability, so future studies should use a larger pool of customer inquiries. Each chatbot was tested with only 15 questions, which may skew the outcomes. Additionally, the selected questions were intentionally chosen for their complexity, potentially biasing the sample toward more articulate customers. Future research should include a broader range of inquiry types to improve generalizability. In terms of transferability, the findings should ideally align with previous studies, but no comparable studies exist due to the novelty of the topic. Although prompting techniques have been studied, this research uniquely evaluates their impact on customer service chatbots. Our conclusion that the CoT-prompted chatbot performed best is provisional, based on 15 selectively chosen inquiries. Therefore, it should be viewed as hypothesis-generating rather than conclusive evidence. Further research with larger sample sizes and more variety in prompts should be conducted.

There are several key limitations. First, data collection took place during the COVID-19 pandemic, which affected the number of inquiries. This context may have influenced chatbot performance, especially in effectiveness and responsiveness. However, aside from the number of inquiries, the content did not differ much from "normal times". Future studies should test the effects of prompting using post-pandemic data and provide a more accurate assessment. Second, some dataset conversations were translated from Finnish to English, potentially losing nuances that could affect the identification of service-quality indicators. Additionally, this study focused on expert opinions rather than customer experiences. Future research should adopt a user-centric perspective, examining factors like perceived usefulness and ease of use. The ethical implications of AI in customer service were out of the scope of this research. Crucial issues, such as hallucinations, data privacy and algorithmic bias, should be explored further. Finally, a reliable model for measuring AI chatbot service quality has yet to be developed. Cultural differences in service preferences further complicate this, and service quality models may not be universally applicable across different business sectors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Tommi Kekäläinen graduated from the Master's Degree Programme in Tourism Marketing and Management in 2024. He currently works as a Marketing Coordinator and Specialist, with a focus on CRM, email marketing, and marketing automation. His professional interests include exploring artificial intelligence in marketing processes and automated customer journeys.

Johanna Heinonen-Kemppi is a university lecturer at the University of Eastern Finland, specialising in digital customer service and the use of AI in designing service experiences and experiential value. Her work explores how caring, co-creation, and technology shape meaningful customer interactions across tourism, education, and service industries.

Juho Pesonen is a professor of tourism business at the University of Eastern Finland, Business School. He works on developing sustainable tourism businesses and tourism destinations. His research focuses especially on the digital transformation of the industry. He is also researching the role of tourism as an experience industry sector.

Erose Sthapit is a Senior Lecturer at the department of marketing, international business and tourism, Manchester Metropolitan University. His research interests include memorable tourism experience and interactive value formation. He has published in high-ranking journals including Journal of Travel Research, Journal of International Contemporary Hospitality Management and Psychology & Marketing.

Brian Garrod is Professor of Marketing in the School of Management at Swansea University. He is author of eight textbook, the most recent being Managing Tourism Attractions, now in its third edition. He has also authored over 80 refereed articles in the field of tourism marketing.

ORCID

Juho Pesonen http://orcid.org/0000-0003-0167-9142
Erose Sthapit http://orcid.org/0000-0002-1650-3900

References

American Express. (2017). Customer service barometer. Business Wire. Retrieved November 4, 2024, from https://www.businesswire.com/news/home/20171215005416/en/WellActually-Americans-Say-Customer-Service-is-Better-Than-Ever

Augustyn, M. M. (1998). The road to quality enhancement in tourism. *International Journal of Contemporary Hospitality Management*, *10*(4), 145–158. https://doi.org/10.1108/09596119810222113

Augustyn, M., & Ho, S. K. (1998). Service quality and tourism. *Journal of Travel Research*, *37*(1), 71–75. https://doi.org/10.1177/004728759803700110

Bacile, T. J. (2020). Digital customer service and customer-to-customer interactions: Investigating the effect of online incivility on customer perceived service climate. *Journal of Service Management*, *31*(3), 441–464. https://doi.org/10.1108/JOSM-11-2018-0363

Berg, J., Buesing, E., Hurst, P., Lai, V., & Mukhopadhyay, S. (2022). *The state of customer care in 2022*. McKinsey & Company.

Blümel, J. H., Zaki, M., & Bohné, T. (2024). Personal touch in digital customer service: A conceptual framework of relational personalisation for conversational AI. *Journal of Service Theory and Practice*, *34*(1), 33–65. https://doi.org/10.1108/JSTP-03-2023-0098

Brazzoli, M. (2023, May 2). *Making sense of textual data: An introduction to quantitative content analysis*. UniAthena. https://uniathena.com/quantitative-content-analysis

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165

Cacic, M. (2023). *5 reasons your business needs a fine-tuned AI model. Entry Point AI*. Retrieved from https://www.entrypointai.com/blog/reasons-your-business-needs-a-fine-tuned-ai-model/

Carvalho, I., & Ivanov, S. (2024). ChatGPT for tourism: Applications, benefits and risks. *Tourism Review*, *79*(2), 290–303. https://doi.org/10.1108/TR-02-2023-0088

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Chen, Q., Gong, Y., Lu, Y., & Tang, J. (2022). Classifying and measuring the service quality of AI chatbot in frontline service. *Journal of Business Research*, *145*, 552–568. https://doi.org/10.1016/j.jbusres.2022.02.088

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. arXiv. https://doi.org/10.48550/arXiv.2209.01390

DeLisi, R., & Michaeli, D. (2021). *Digital customer service: Transforming customer experience for an on-screen world*. Wiley.

De Keyser, A., Köcher, S., Alkire, L., Verbeeck, C., & Kandampully, J. (2019). Frontline service technology infusion: Conceptual archetypes and future research directions. *Journal of Service Management*, *30*(1), 156–183. https://doi.org/10.1108/JOSM-03-2018-0082

Dickson, T. J., & Huyton, J. (2008). Customer service, employee welfare and snowsports tourism in Australia. *International Journal of Contemporary Hospitality Management*, *20*(2), 199–214. https://doi.org/10.1108/09596110810852177

Dogra, J. (2024). ChatGPT and its significance in tourism sector: Current scenarios and future roadmaps. *Journal of Multidisciplinary Academic Tourism*, *9*(3), 191–199. https://doi.org/10.31822/jomat.2024-9-3-191

Dogru, T., (Dr. True), Line, N., Mody, M., Hanks, L., Abbott, J., Acikgoz, F., Assaf, A., Bakir, S., Berbekova, A., Bilgihan, A., Dalton, A., Erkmen, E., Geronasso, M., Gomez, D., Graves, S., Iskender, A., Ivanov, S., Kizildag, M., Lee, M., Lee, W., Luckett, J., … Zhang, T. (2025). Generative artificial intelligence in the hospitality and tourism industry: Developing a framework for future research. *Journal of Hospitality & Tourism Research*, *49*(2), 235–253. https://doi.org/10.1177/10963480231188663

Dwivedi, Y. K., Pandey, N., Currie, W., & Micu, A. (2024). Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: Practices, challenges and research agenda. *International Journal of Contemporary Hospitality Management*, *36*(1), 1–12. https://doi.org/10.1108/IJCHM-05-2023-0686

Efrat, A., & Levy, O. (2020). The Turking Test: Can language models understand instructions? arXiv. https://doi.org/10.48550/arXiv.2010.11982

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gentile, C., Spiller, N., & Noci, G. (2007). How to sustain the customer experience: An overview of experience components that co-create value with the customer. *European Management Journal*, *25*(5), 395–410. https://doi.org/10.1016/j.emj.2007.08.005

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Graef, R., Klier, M., Kluge, K., & Zolitschka, J. F. (2021). Human-machine collaboration in online customer service: A long-term feedback-based approach. *Electronic Markets*, *31*(2), 319–341. https://doi.org/10.1007/s12525-020-00420-9

Grönroos, C. (1984). A service quality model and its marketing implications. *European Journal of Marketing*, *18*(4), 36–44. https://doi.org/10.1108/EUM0000000004784

Guttentag, D. A., Litvin, S. W., & Teixeira, R. (2025). Human vs. AI: Can ChatGPT improve tourism product descriptions? *Current Issues in Tourism*, *28*(22), 3601–3619. https://doi.org/10.1080/13683500.2024.2402563

Heinonen, J., & Pesonen, J (2022). Identifying the elements of great online customer encounters. In J. L. Stienmetz., et al. (Eds.), *Information and communication technologies in tourism 2022* (pp. 271–281). Springer. https://doi.org/10.1007/978-3-030-94751-4_24

Hennigan, L. (2024). *How to improve customer service in 2024, Forbes Advisor*. Retrieved November 4, 2024, from https://www.forbes.com/advisor/business/how-improve-customer-service/

Henrickson, L., & Meroño-Peñuela, A. (2025). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & Society*, *40*(2), 903–918. https://doi.org/10.1007/s00146-023-01752-8

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing*, *88*(5), 1–23. https://doi.org/10.1177/00222429231224748

Hudson, S., & Hudson, L. (2013). *Customer service in tourism and hospitality*. Goodfellow Publishers.

Hsu, C.-L., & Lin, J. C.-C. (2023). Understanding the user satisfaction and loyalty of customer service chatbots. *Journal of Retailing and Consumer Services*, *71*, 103211. https://doi.org/10.1016/j.jretconser.2022.103211

Hyken, S. (2024). *The ACA study: The state of customer service and achieving customer amazement*. Shepard Presentations LLC.

Janssen, A., Grützner, L., & Breitner, M. H. (2021). *Why do chatbots fail? A critical success factors analysis* [Paper presentation]. Forty-Second International Conference on Information Systems.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022)*.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2023). Better zero-shot reasoning with role-play prompting. arXiv. https://doi.org/10.48550/arXiv.2308.07702

Lee, S. M., & Lee, D. (2020). "Untact": A new customer service strategy in the digital age. *Service Business*, *14*(1), 1–22. https://doi.org/10.1007/s11628-019-00408-2

Lohrbeer, T. (2023, December 4). *Retrieval-augmented prompting: Enabling prompt switching in GPTs. Medium*. Retrieved November 16, 2024, from https://medium.com/@FastFedora/retrieval-augmented-prompting-enabling-prompt-switching-in-gpts-521821840afa

Li, L., Zhang, Y., & Chen, L. (2023). *Prompt distillation for efficient LLM-based recommendation* [Paper presentation]. Proceedings of the 2023 ACM International Conference on Information and Knowledge Management (CIKM '23) (pp. 1348–1357). ACM. https://doi.org/10.1145/3583780.3615017

Linneberg, M. S., & Korsgaard, S. (2019). Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal*, *19*(3), 259–270. https://doi.org/10.1108/QRJ-12-2018-0012

Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., & Wu, B. (2023). Fairness-guided few-shot prompting for large language models. In *37th Conference on Neural Information Processing Systems (NeurIPS, 2023). Tianjin University; Tencent AI Lab; Institute of High Performance Computing (IHPC)*. Agency for Science, Technology and Research.

MacPhail, C., Khoza, N., Abler, L., & Ranganathan, M. (2016). Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research*, *16*(2), 198–212. https://doi.org/10.1177/1468794115577012

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://arxiv.org/abs/2303.08774

OpenAI. (2024, August 8). *GPT-4o system card*. Retrieved October 11, 2024, from https://openai.com/index/gpt-4o-system-card/

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41–50. https://doi.org/10.1177/002224298504900403

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *64*(1), 12–40.

Parasuraman, A., Zeithaml, V. A., & Malhotra, A. (2005). E-S-QUAL: A multiple-item scale for assessing electronic service quality. *Journal of Service Research*, *7*(3), 213–233. https://doi.org/10.1177/1094670504271156

Paulus, T. M., Pope, E. M., Woolf, N., & Silver, C. (2019). "It will be very helpful once I understand ATLAS.ti": Teaching ATLAS.ti using the Five-Level QDA method. *International Journal of Social Research Methodology*, *22*(1), 1–18. https://doi.org/10.1080/13645579.2018.151066

Rospigliosi, P. A. (2023). Artificial intelligence in teaching and learning: What questions should we ask of Chatgpt? *Interactive Learning Environments*, *31*(1), 1–3. https://doi.org/10.1080/10494820.2023.218019
Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development*, *52*(1), 5–18. https://doi.org/10.1007/BF02504769
Van Glabbeek, R. J. (2001). The linear time-branching time spectrum I: The semantics of concrete, sequential processes. In J. A. Bergstra, A. Ponse, & S. A. Smolka (Eds.), *Handbook of process algebra* (pp. 3–99). Elsevier.
Wang, P. Q. (2025). Personalizing guest experience with generative AI in the hotel industry: There's more to it than meets a Kiwi's eye. *Current Issues in Tourism*, *28*(4), 527–544. https://doi.org/10.1080/13683500.2023.2300030
Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., Zhang, M., Zhang, Z., Ouyang, W., Xu, K., Huang, S. W., Fu, J., & Peng, J. (2023). RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2310.00746
Wang, S., & Zhang, H. (2025). Artificial intelligence, digital employees and sustainable innovation in online retail: The mediating role of ambidextrous green innovation and the moderating role of ethical anxiety. *Journal of Retailing and Consumer Services*, *84*, 104235. https://doi.org/10.1016/j.jretconser.2025.104235
Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations (ICLR). Google Research*. https://iclr.cc/
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models* [Paper presentation]. 36th Conference on Neural Information Processing Systems (NeurIPS, 2022). Google Research, Brain Team.
Wong, A. W. O. M., Dean, A. M., & White, C. J. (1999). Analysing service quality in the hospitality industry. *Managing Service Quality: An International Journal*, *9*(2), 136–143. https://doi.org/10.1108/09604529910257920
Wu, T. Y., He, S. Z., Liu, J. P., Sun, S. Q., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo, and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618
Zamfrescu-Pereira, J. D., Wong, R., Hartmann, B., & Yang, Q. (2023). *Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts* [Paper presentation]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023 (pp. 1–21). ACM. https://doi.org/10.1145/3544548.3581388
Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2023). When "A helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. arXiv. https://doi.org/10.48550/arXiv.2311.10054
Zdrok, O. (2024, March 19). *Master the prompt: 7 contrasts between zero-shot and few-shot prompting*. Shelfio. https://shelf.io/blog/zero-shot-and-few-shot-prompting/

Appendix 1. Chatbot prompting

Chatbot A (zero-shot)

Instruction: You are a Visit Helsinki customer service chatbot. Respond politely and helpfully to customer inquiries in a professional tone using the same language as the question.

Respond to the question:

Chatbot B (few-shot)

Instruction: You are a Visit Helsinki customer service chatbot. Respond politely and helpfully to customer inquiries in a professional tone using the same language as the question.

Example 1: **Input:** Hei, haluaisin löytää jonkun lounasristeilyn helsingin kauppatorilta, 2–3h tämän viikon torstaina. Osaatteko neuvoa miltä sivustolta löytäisin? Meitä on neljä henkilöä.

Output: Hei! Kiitos kysymyksestä! Pieni hetki.

Esimerkiksi Royal Linella on lounasristeilyitä, jotka lähtevät Kauppatorilta. https://www.royalline.fi/Lounas-ja-illallisristeilyt

Nämä ovat kestoltaan tosin vain puolitoista tuntia.

Usein laivaan voi kuitenkin saapua jo noin puoli tuntia ennen lähtöä, mutta tämä kannattaa vielä varmistaa laivayhtiöltä.

Input: Kiitos, sinne siis! Ei haittaa, keksimme muutakin ohjelmaa siinä tapauksessa. Tuossa ajassa ehtii myös syödä rauhassa.

Output: Eipä kestä. Mukavaa viikkoa ja risteilyä!

Example 2: **Input:** Hei. Mietimme juuri että onko Helsingissä nyt kesällä tapahtumia? Kaikki musiikkitapahtumat kiinnostavat.

Output: Hei, kyllä niitä jonkin verran on, vaikka toki vähemmän kuin normaalina kesänä

Tapahtumista löytyy parhaiten tietoa myhelsnki.fi -sivustolta tapahtumahaku -toiminnosta tai esim. stadissa.fi -sivustolta

Input: Hei kiitos sinulle ja ihanaa viikonloppua!:)

Output: Kiitos, sitä samaa teille!

Example 3: **Input:** Hei! We want to go to Seurasaari but not sure the best parking options. Any suggestions?

Output: Hello! It is a beautiful place. Just a moment, please. I will send the information.

Input: Yes, it is. Thanks

Output: At the bottom of this Seurasaari web page you can find parking information: https://www.kansallismuseo.fi/en/seurasaarenulkomuseo/info

Input: Got it. Thanks.

Output: Great. Have a nice day!:)

Input: I am trying to find this in google maps: Next to the sportsfield of Meilahti (about 1 km) is a large free parking area.

I don't see it.

Output: Just a moment, I will look for it.

Google Maps

Input: Thanks

Output: And this is the sport center of Meilahti

Google Maps

Input: Great

Output: Good:) The parking is next to the sport center.

Respond to the next question:

Chatbot C (chain-of-thought):

Instruction: You are a Visit Helsinki customer service chatbot. Respond politely and helpfully to customer inquiries in a professional tone using the same language as the question.

Example: Input: "Hei, haluaisin löytää jonkun lounasristeilyn Helsingin Kauppatorilta, 2–3h tämän viikon torstaina. Osaatteko neuvoa miltä sivustolta löytäisin? Meitä on neljä henkilöä."

Output:

The customer is asking about a specific service: for example a lunch cruise from Helsinki's Kauppatori for 2–3 h this Thursday. First, greet the customer and thank for their message. Tell them it will take a while to find an answer.

Next, Look up or confirm available lunch cruises that fit the customer's requirements, paying attention to the duration and departure point.

Third, Provide a suggestion based on the search results, including a link to the service provider's website for convenience.

Add any helpful details, such as boarding times or other practical information that the customer might find useful, especially if the options don't exactly fit their original request.

Confirm that the information meets the customer's needs and respond to any follow-up comments or concerns.