Generative AI chatbot prompting for excellent customer service in tourism

Tommi Kekäläinen (M.Sc)
University of Eastern Finland
Business School
Yliopistonkatu 2, 80101 Joensuu, Finland
tommi.kek@outlook.com

Erose Sthapit (PhD)
Corresponding Author
Senior Lecturer in Tourism
Department of Marketing, Retail and Tourism
Manchester Metropolitan University
All Saints, All Saints Building, Manchester M15 6BH
e.sthapit@mmu.ac.uk
ORCID: 0000-0002-1650-3900

Brian Garrod (PhD)
Professor in Marketing
School of Management
Swansea University
Bay Campus, Fabian Way, Swansea,
SA1 8EN
brian.garrod@swansea.ac.uk
ORCID: 0000-0002-5468-6816

Johanna Heinonen-Kemppi (PhD)
University lecturer
University of Eastern Finland
Business School, Yliopistonkatu 2, 80101 Joensuu, Finland
LAB University of Applied Sciences, Yliopistonkatu 36, 53850 Lappeenranta, Finland
johanna.heinonen@lab.fi

Juho Pesonen (PhD)
Professor of Tourism Business,
University of Eastern Finland,
Business School, Yliopistonkatu 2, 80101 Joensuu, Finland
Juho.pesonen@uef.fi
ORCID: 0000-0003-0167-9142

Generative AI chatbot prompting for excellent customer service in tourism

Abstract

This paper explores the impact of prompting on the performance and service quality of artificial intelligence customer-service chatbots, and the connection between the prompting process and the elements of service quality. 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 an important influence on the quality of online customer service that AI chatbots can provide.

Keywords: artificial intelligence; chatbots; customer service; prompt engineering; service quality

1 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 have many positive features, including the ability to converse naturally and provide personalised recommendations (Dogru et al., 2023). This means that AI chatbots are capable of delivering high levels of service quality, which can have positive effects on purchase intentions, customer satisfaction and brand engagement (Blumel et al., 2024).

Chatbots have, however, been found to have a high failure rate, meaning that they do not properly understand the customer's request and fail to provide an optimal response. The ability of Chatbots to read customers' emotions correctly and respond accordingly is also vital in customer-service contexts (Huang & Rust, 2024). When chatbots fail, this can lead to a reduction in customer trust and negative word-of-mouth, which can be reflected in falling sales and reduced 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 being one of the potential barriers. 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 optimise its outputs (Henrickson & Meroño-Peñuela, 2023). 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 brings forth novel issues such as promises for saving costs and increasing work efficiency, but also concerns about how AI can handle complex interactions in a culturally suitable way (Wang, 2025).

The purpose of this study is, therefore, to investigate the impact of different prompting techniques on the quality of chatbot responses with respect to various dimensions of customer service in the tourism context. In doing so, the study aims to understand how chatbots can contribute to enhancing customer service and overall customer satisfaction for tourism businesses. As will be explained in the following section of the paper, the focus will be on how prompting might best be employed 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 the delivery of services, and this brings with it the need for the 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 customer services 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 many potential applications in the tourism context (Carvalho & Ivanov, 2024).

2 Literature Review

2.1 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 can be defined as the interaction between the customer and the service provider. This has traditionally been done face-to-face (Hudson & Hudson, 2021), but increasingly the relationship is being digitalised, 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 that customer service is more important than ever, and 87% of customers feel that good customer service increases trust in a company or brand. Meanwhile, 33% of customers were willing to change service providers immediately after a bad customer-service experience (American Express, 2017).

2.2 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 organisation, at all touch points throughout their customer journey. Gentile et al. (2007), indeed, argue that the traditional drivers of competitiveness, such as price and product quality, are no longer enough for most companies: maintaining competitive advantage requires companies to provide excellent customer service. AI is changing digital content creation, but its impact on competitive advantage is still under scrutiny (Guttentag et al., 2024).

While service quality is recognised to play an important role in determining the overall customer experience, it has proven more difficult to measure the quality of a service than that of a physical product. It is possible to examine goods tangibly, based on their various characteristics such as style, colour, or feel. Services, however, lack this tangible dimension, making it much more difficult to know how to measure it (Parasuraman et al., 1988; Grönroos, 1984). One of the best-known contributions to the understanding of service quality is the SERVQUAL model introduced by Parasuraman et al. (1985). The SERVQUAL concept suggests that customers assess service quality across five specific 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), fulfilment (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 are increasingly moving online and AI is being used to deliver these services, 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' (AISQ). The model recognises that the characteristics of AI chatbot services differ significantly from human-provided services. For example, an AI chatbot can hold a much more significant amount of 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, Human-like, Continuous improvement, Personalisation, Cultural adaptation, and Efficiency.

2.3 AI in customer service

2.3.1 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 has, however, disrupted the status quo, and digital technologies are now being integrated into all areas of business. Customer service is no exception and is increasingly being delivered digitally. Customer preferences and shopping behaviour have also shifted significantly towards online services, justifying the need for investing in different forms of digital customer services (Lee et al., 2019; Bacile, 2020; DeLisi & Michaeli, 2021). The digitalisation of customer service has been described as a 'win-win-win-win': customers get better, faster service, 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 how to transform traditional, analogue customer-to-customer service representative interactions – 'moments of truth' – to a digital format: meeting customers where they are (DeLisi & Michaeli, 2021). The platforms where these encounters can take place may be in social media (for instance, on 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 analyse data from various customer service situations to provide feedback to the human customer service agent. Second is conversational coaching, which uses the data collected from different customer-service situations to make suggestions for improvement or even to supplement the human customer-service agent's messages and make recommendations. These first two types of service encounters are known as 'AI-supported service encounters.' Third are chatbots, which are designed to replace human agents, responding 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 still share the workload in practice: the chatbot can handle simple questions and problems, and the conversation is passed on to the human customer service agent if problems 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 and 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) analysed 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 apologising), and personalisation (providing tailor-made messages to each customer and identifying their personal needs).

2.3.2 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 do, however, rely on text inputs, which 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 using artificial intelligence is ChatGPT. It is a machine-learning software utilising the GPT ('Generative Pre-trained Transformer') 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 leading to reduced costs and a reduced need to employ staff. There will also be benefits for users, as services powered by ChatGPT will open up opportunities for efficient, fast service around the clock (Gursoy et al., 2023).

ChatGPT is expected to revolutionise many different business sectors, and the tourism industry is no exception. One of the main uses of ChatGPT in tourism is the personalised advice it can provide. It can, for example, make suggestions for a travel itinerary, including places to visit, eat, and stay – all taking into account the personal preferences or constraints of the customer (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 be able to reach human-level performance in 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 utilising 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 personalised 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). Optimising the prompts with more context and examples leads to better and accurate responses. Nevertheless, creating effective prompts can be a challenging task. 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 prompting method that involves specific instructions or questions to the LLM, without giving any specific background or dataset. The LLM then refers to its knowledge base and answers based on the prompt (Zdrok, 2024). When using this prompting technique, therefore, the prompt needs to be as precise and detailed as possible, so that there is no room for misinterpretation by the LLM. Henrickson and Meroño-Peñuela (2023) found in their study that adding 'zero-knowledge' to the prompt can improve the accuracy of zero-shot prompting. This means giving the question a little more context by providing instructions that describe the task (for example: "let us think step by step"). This can be particularly helpful in prompts that involve counting, as language models cannot think abstractly (Kojima et al., 2022).

'Few-shot' prompting is a step up from zero-shot prompting. Its purpose is to condition the language model by giving it some examples (Brown et al., 2020). As the LLM has already been trained with large datasets, these specific examples help it to approach the task from the right angle more efficiently (Zdrok, 2024). Few-shot prompting has been shown to perform relatively well. However, it tends to suffer from instability due to different variations in the examples used, their order, and the different formats of prompts. (Ma et al., 2023).

Another prompting tool that has been found to improve language model inference is the chain-of-thought technique (CoT). This involves breaking down a problem into smaller parts and solving them sequentially before coming to a final answer, allowing additional steps to be used in more demanding problems. This also makes troubleshooting easier, as it is possible to trace the step at which a wrong inference was made (Wei et al., 2022).

There are thus several different prompting techniques (see Table 1). Often, good output results require the use of trial-and-error to test the different techniques (Zamfrescu-Pereira et al., 2023).

Table 1

2.3.3 Prompting as a process

Prompting thus involves the formulation and presentation of a command or instruction to a language model, the aim being to improve the 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 behaviour of a system, whether it is a machine, a protocol or, in the present case, a chatbot. Process theory involves two main activities: modelling (representing processes in their own system language) and verification (proving statements about the process, such as whether the behaviour of the system is as intended). In relation to the prompting process, modelling involves the creation of prompts, while verification focuses on testing prompting methods and comparing the 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 analysed sequentially. Each prompt leads to a particular output or response, which means that different prompting methods can be viewed as different process 'semantics.'

3 Method

This paper aimed to examine which of a chosen set of prompting methods are the most effective in helping to improve 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 personalisation. The study used a mixed deductive/inductive approach. This involved coding and categorising data in pre-assigned categories, with new subcategories emerging as the data were analysed.

3.1 Data collection

Based on the results of studies in the general context (Brown et al., 2020; Zdrok, 2024), three different 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 Organisation based in Helsinki, Finland) customer service chat conversations in June and July 2020, using these techniques.

From a total of 123 chat conversations, 15 different customer questions were chosen by purposive sampling, based on those that had produced long, detailed and, in the researchers' opinion, high-quality answers. To ensure a variety of difficulty levels, we selected five questions each from three categories: simple, moderately complex, and complex. Out of a total of 41 questions in English and 83 in Finnish, our sample included four questions 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 different nature and complexity than typical tourist contexts.

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 in this paper 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.

3.2 Data analysis

Quantitative content analysis (QCA) was chosen as the research method. Despite its name, it is classified as a qualitative research method and is useful for identifying trends or patterns in large textual data sets (Brazzoli, 2023). QCA is known as a systematic and objective procedure for describing communication, including segmenting the data into units and categories and later creating summaries of these categories (Rourke & Anderson, 2004). In QCA, coding involves categorising the data, i.e., using codes to summarise and label the data (which could be a word or phrase), into specific notes or memos. This process allows further analysis or visualisation of the data (Linneberg & Korsgaard, 2019). The aim was to find differences in responses between different chatbots and to identify which chatbot had the best performance 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 personalisation. 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.

Figure 1

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 through "Proactive communication" and "Informing of delays," alongside accuracy, relevance, and completeness of information, 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 different chatbots were uploaded to Atlas.ti, which is a software package commonly used for qualitative data analysis (Paulus et al., 2017). Data collected from chatbot responses were coded by carefully reading through the data phrase by phrase and coding each line according to one of the categories 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 allowed evaluation of the occurrence of different codes and a comparison between the performance of the three chatbots. Coding was first performed by a single researcher. Then, an independent coder performed another round of coding. We followed intercoder reliability process guidelines based on MacPhail et al. (2016). Intercoder reliability was assessed retrospectively on 10 double-coded transcripts (22% of sample) using Cohen's Kappa. Overall agreement was substantial ($\kappa = 0.78$, 95% CI: 0.72-0.85, $p < .001$). 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. Analysis revealed this reflected a methodological difference rather than random error. The second coder applied overlapping quality indicators inclusively, while the other coded selectively. Qualitative examination confirmed that both coders used codes from the same thematic groups and independently identified identical performance rankings (Table 4), supporting construct validity despite coding methodology differences. The results reported in this paper are based on the first coder analysis, who analysed the complete data set.

4 Results

This section of the paper presents the results in terms of the performance of the three chatbots in the four dimensions of good customer service. A total of 15 questions were tested with each of the three chatbots, resulting in a total of 45 responses. The topics of the questions can be roughly divided into four main groups: practical information and local regulations (four questions), activities and attractions (four questions), local venue recommendations (four questions), and transportation and accommodation (four questions).

4.1 *Chatbot Performance*

Following the coding process, the frequency of each code in the responses of the different chatbots was quantified. From the quantities of codes, the performance of the chatbot in different dimensions of customer service can be inferred. The detailed breakdown of this performance is presented in Table 2, allowing for comparison of the chatbots' abilities in delivering quality customer service. In the table, "Gr" refers to the so-called 'groundedness' of codes (i.e., the number of quotations covered by a code).

Table 2

'Personalisation' 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 Apologising. ‘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 actively kept the customer updated.

4.2 Classification of coded data

Table 1 shows the performance of different chatbots in different dimensions of customer service. The quantities of quotes matching the codes were then classified according to qualitative descriptions of chatbot performance to simplify data interpretation. The performance score of the chatbots ranged from 0 to 31, based on which five different categories were developed to reflect different levels of performance. This made it easier to describe the strengths and weaknesses of each chatbot: for example, which was best at giving personalised responses or which was most polite (Table 3).

Table 3

4.3 Findings

4.3.1 Summary of chatbot performance in customer-service dimensions

By combining the results from Table 2 and the code range descriptions from Table 3, the results of the chatbot responses were put into an easy-to-read format (Table 4). The results show that Chatbot C performed significantly better than Chatbot A and Chatbot B. The analysis of Chatbots A and B resulted in almost the same number of coded quotes (a total of 51 for Chatbot A and 54 for Chatbot B), varying slightly in different areas of the customer service dimensions, while Chatbot C scored a significantly higher total with 95 coded quotes.

Table 4

4.3.2 Personalisation

Chatbot C was the best performer in the ‘Personalisation’ dimension, receiving a rating of ‘Excellent’ (total of 31 coded quotes), reflecting its ability to produce well-tailored answers to questions. Overall, Chatbot A performed reasonably well, receiving a ‘Good’ overall (total of 12 coded quotes) rating in ‘Personalisation,’ while Chatbot B’s performance was only ‘Moderate’ (total of 8 coded quotes). However, the differences between these two chatbots were relatively small, with the most noticeable difference being in the ‘Tailored response’ subcategory, where Chatbot A scored five coded quotes and Chatbot B scored only one.

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’:

"... For transportation, you can consider the ***HSL 3-day ticket****, 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).

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 poorly to support their answers, with both receiving 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 was able to find places that were especially suitable for children.

4.3.3 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 quotes, along with Chatbot B, which received 25. Chatbot A, meanwhile, received a total of only 12 coded quotes in this dimension.

None of the chatbots apologised in their response for any of the 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 quotes for 'Apologising.' Chatbot C excelled in both 'Thanking' and 'Greeting,' and scored 15 coded quotes, thanking the user in each response. Chatbot B similarly greeted in each response, receiving 15 coded quotes 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 a total of only three coded quotes for 'Thanking' and nine for 'Greeting,' being the weakest in this category.

4.3.4 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 24 coded quotes, Chatbot B receiving 28, and Chatbot C receiving 26. This resulted in a rating of 'Excellent' for B and C, and 'Very Good' for Chatbot A.

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 a total of 11 coded quotes in this subcategory. However, Chatbot B (10 coded quotes) and Chatbot C (eight coded quotes) were not far behind, both often giving similar tips in their responses to Chatbot A.

'Fulfilling expectation' also received a very similar number of coded quotes across all three Chatbots. In contrast to the study by Heinonen and Pesonen (2022), where the outcome of the discussion 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 13 coded questions, while Chatbot B received 16, 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.

4.3.5 Effectiveness

In terms of efficiency, Chatbot C performed the best with 11 coded quotes, receiving a rating 'Good,' followed by Chatbot A, which had a total of seven coded quotes with a rating 'Moderate,' and Chatbot B with four coded quotes, resulting in a 'Poor' rating. The chatbots generally did not inform the user about delays in responses, with Chatbots B and C each doing so only once, while Chatbot A did not do this 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 analysing the chatbot responses, the second identified sub-category was 'Proactive communication,' which refers to how the chatbot kept updated with its responses. 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 10 coded quotes, while Chatbot A received seven coded quotes and Chatbot B received three.

5 Discussion

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

5.1.1 Prompting process

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

Table 5

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

The next phase of the prompting process is modelling, 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 viewed as central activities of the process theory (Van Glabbeek, 2001). This part of the process relates to functional quality, as Grönroos (1984) defines it, which refers to the service-quality dimension that corresponds to how service quality is perceived.

The prompting process also involves troubleshooting and fixing any errors that occur, to keep the outputs of the chatbot at a consistently high quality. This aligns with the process theory element known as 'failure tracing,' which involves observing the process to understand why the system has failed to succeed under certain conditions (Van Glabbeek, 2001). Regarding service-quality theories, this part of the process relates to the AISQ theory dimension 'continuous improvement' (Chen et al., 2022), enhancing the system stability and improving its capabilities.

The final phase of the prompting process relates to the output of the prompt and how it is presented. It is the tangible and most visible part of the process, shaped by the various actions and transitions throughout, which is also a key dimension of process theory. This relates to how processes evolve: by performing a certain action, a process transitions into another process, thus affecting the result (Van Glabbeek, 2001). The chatbot prompting process thus embodies the system's ability to transition from a prompt into a concrete and 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).

5.1.2 Prompting methods

In the foregoing analysis, Chatbot C, which applied the CoT prompting method, outperformed the other two chatbots in almost every dimension of excellent online customer service. This suggests that a chatbot prompted by a CoT method delivers the highest quality of customer service across most dimensions of online customer service and may therefore be seen as the best prompting method in this context. Chatbot C delivered a very high level of personalisation, responding in a specific and tailored manner to the questions and demonstrating context awareness. Chatbot B (which used few-shot prompting) performed reasonably well, being the best in terms of responsiveness (where all chatbots performed well) but lagging 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 behaviour 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.

Table 6

Overall, the CoT method performed best, but the zero-shot method also had advantages in terms of generating concise and 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.

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 make responses excessively lengthy for simple queries. Therefore, prompt engineers should choose methods based on their strengths while being aware of potential weaknesses.

Ethical considerations in AI-powered customer service are crucial. Customers often experience discomfort or distrust when interacting with non-human agents, which can be described as ethical anxiety (Wang & Zhang, 2025). Such concerns may impact the relationship between perceived service quality and customer satisfaction, highlighting that even excellent technical responses may not ensure a positive experience.

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 miscommunication risks. The few-shot method's effectiveness relies on high-quality examples. Overall, prompt engineering serves not only to enhance service quality but also to minimise the risk of critical chatbot failures.

6. Conclusions

6.1. Theoretical and methodological contribution

The main theoretical contribution of this paper relates to how the use of different prompting methods can improve the results of AI-powered LLMs. The results suggest that the use of specific prompting methods can lead to better results in different dimensions of excellent customer service and that there is a link between prompt structure and chatbot behaviour. In short, chatbots can provide better customer service if they are suitably promoted.

The results also contribute to process theory. Prompting can be viewed as a structured process that leads to quality chatbot responses. By combining the process of prompting with the elements of service quality, insights can be gained into how different steps in the prompting process can affect the quality of chatbot responses.

This paper also contributes methodologically. Prompting customer service chatbots and testing them with fundamental questions collected from customer conversations is a relatively new way of designing and evaluating the performance of chatbots. Testing with real questions ensures that the results reflect authentic customer-service situations, and the differences in the results show which approach can be more effective in practice. This study thus contributes to the development of realistic and user-centred AI chatbots in the future.

6.2. Managerial implications

With respect to managerial implications, one of the most important contributions of this study is the increased knowledge of the range of possibilities regarding how to improve the responses of AI-driven customer-service chatbots. By demonstrating how specific prompting methods can affect chatbot effectiveness, responsiveness, personalisation and politeness, organisations can leverage different programming techniques for their own customer-service needs.

Indeed, this study demonstrates how it is possible to tailor or fine-tune chatbots according to the specific needs of companies through a process of prompting. Managers can thus utilise prompting to meet the organisation's business objectives. For example, if a company had set itself the goal of conveying an image of friendliness, the customer-service chatbot could be prompted to be as friendly as possible. If the company wanted to convey a different image, such as formality, the chatbot could be programmed differently to achieve this aim. Choosing a particular prompting method can also allow the organisation to shape what kind of responses are produced. For example, the CoT prompting method produced the longest and most detailed responses in this study, which is likely to be most suitable for detailed or complex inquiries. The zero-shot prompting method, meanwhile, produced shorter and more straightforward responses, which would be more suitable for shorter and simpler inquiries.

This study highlights that the effectiveness of chatbots depends not only on the prompting strategies but also on the organisational ecosystem surrounding them. Prompt engineering is just one aspect of a broader AI strategy that requires strong support, continuous training, and the development of employees' technological skills. Effective integration helps organisations manage AI systems while enhancing employee performance and creativity. Recent research in hospitality (Wang & Zhang, 2025) shows that organisational support is closely linked to successful AI implementation.

To make this practical, managers can use a simple framework when choosing a prompting method. First, they should ask: What is the primary service objective? For transactional efficiency (like frequently asked questions or 'FAQs'), zero-shot prompting is best for concise responses. For balanced interaction (moderately complex tasks), few-shot prompting is appropriate with carefully selected examples. For complex problem-solving or building relationships (like itinerary planning), the chain-of-thought method is recommended for richer, personalised responses. This framework helps align prompting strategies with customer service goals. This study highlights the importance of understanding the role and effects of prompting on all business levels.

More generally, the implementation of chatbots can contribute to a company's efficiency by automating labour-intensive tasks. Cacic (2023) explains that by choosing the appropriate prompting method and fine-tuning it to fit the company's specific needs, it is possible to handle complex cases automatically, saving time spent by a human customer-service agent. When a chatbot has not been fine-tuned, the human customer-service agent may find themselves being 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 be reflected in customer satisfaction.

6.3. Limitations and suggestions for future research

This study has some limitations that can affect the quality of the results. The limited size of the sample may have compromised the reliability of the findings. To achieve more reliable results, future studies should employ a larger sample of customer inquiries. Each chatbot was tested with only 15 different questions, which could have affected the results. A wider range of prompting methods could also have been used to improve the reliability of the results.

In terms of transferability, ideally, the results of the study should be broadly comparable with those of previous studies. As this is a relatively new topic, however, no examples of previous similar studies were available for this purpose. Prompting techniques have been studied in the past, but not in the context of evaluating the quality of customer service chatbots.

This study has several limitations. First, intercoder reliability was assessed retrospectively rather than iteratively. While overall agreement was substantial ($\kappa = 0.78$), systematic differences in coding approach, such as response quality categories being mutually exclusive vs overlapping, resulted in low Kappa for some subcategories despite qualitative agreement on code themes. Despite these methodological differences, both coders independently reached identical substantive conclusions (Table 4), supporting the robustness of findings.

Second, data collection occurred during the summer of 2020, at the peak of the COVID-19 pandemic. This context may not accurately reflect typical customer service inquiries. Future studies should use post-pandemic data for a better assessment.

Additionally, this study only evaluated expert opinions on chatbot outputs, ignoring the customer experience. Future research should adopt a user-centric perspective, looking at factors like perceived usefulness, ease of use, interactivity, and immersion within an extended TAM.

The ethical dimensions of AI-powered customer service were also overlooked. Concerns about authenticity, fairness, data privacy, and algorithmic bias are critical and should be explored in future studies.

Finally, as the field of generative AI is rapidly evolving, this study's focus on the GPT-4 model represents only a snapshot of current capabilities, as newer models may change effective prompting methods. However, prompting will most likely be a critical part of interactions with these AI chatbots, as prompting instructs the chatbot about user needs.

Finally, it is important to recognise that a fully valid and reliable model for measuring service quality has yet to be developed, let alone in the context of AI chatbot service quality. Shortcomings remain, for example, in respect of the tendency for people of different cultures to have their own specific preferences and tolerances toward customer-service provision and service failure. A model developed in one culture may therefore not work well when applied to another. It has also been noted that service-quality models are not universally applicable to every area of business, and those that work well in the context of tourism might not work equally well in other business contexts (Seth et al., 2005).

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.

Blumel, 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), Article 39. 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.

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: Hoboken.

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., 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., & Zhang, T. (2023). Generative artificial intelligence in the hospitality and tourism industry: Developing a framework for future research. *Journal of Hospitality & Tourism Research 49*(2), 235–225. 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.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*(10), 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. (2024). Human vs. AI: Can ChatGPT improve tourism product descriptions? *Current Issues in Tourism*, 1–19. 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. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & Society*. 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. 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. (2019). "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. In 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.

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. (2017). "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

Rich, E. (1985). Artificial intelligence and the humanities. *Computers and the Humanities, 19*(2), 117–122.

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

Seth, N., Deshmukh, S. G., & Vrat, P. (2005). Service quality models: A review. *International Journal of Quality & Reliability Management, 22*(9), 913–949. https://doi.org/10.1108/02656710510625211

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., Sun, Z., Li, M., Zhang, H., & Metwally, A. H. S. (2024). Leveraging TikTok for active learning in management education: An extended technology acceptance model approach. *International Journal of Management Education, 22*(3), 101009. https://doi.org/10.1016/j.ijme.2024.101009

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. In 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. http://dx.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 Automatika Sinica, 10*(5), 1122–1136.

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. In 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. Shelf.io. https://shelf.io/blog/zero-shot-and-few-shot-prompting/