RESPONSible Service: Evaluating Language Style in LLM-Generated Customer Service Interactions
Bigontina, Nicolo' de

RESPONSible Service: Evaluating Language Style in LLM-Generated Customer Service Interactions

Nicolò de Bigontina

Leiden University
MA Linguistics: Computational Linguistics – Graduation Thesis

Supervisor: Prof.dr. Stephan A. Raaijmakers
Second reader: Prof.dr. Carole P.A. Tiberius

June 27, 2025

Abstract

Language style plays a crucial role in effective customer service, and the growing use of AI chatbots in this field underscores the need to assess their ability to recognize and reproduce stylistic variation across different scenarios. However, there is a lack of publicly available datasets specifically focused on language style in customer service. To address this gap, I created RESPONSible Service, a synthetic dataset of 4,000 GPT-4o-generated customer service interactions. The dataset is divided into four subsets, each targeting a specific communication style: clarity, friendliness, empathy, and politeness. In the absence of a human-authored ground truth, I evaluated the dataset using a range of NLP metrics and compared its stylistic features to linguistic patterns identified in prior research on human customer service communication. Results show that GPT-4o generally aligns with human communication trends, although it occasionally shows model-specific deviations and inconsistencies in context-sensitive generation. To assess the dataset's usability for text generation, I fine-tuned a smaller model (LLaMA 3.2-1B-Instruct) using in-context learning, supervised fine-tuning, and direct preference optimization. Outputs were evaluated using BERTScore, ROUGE-L, and manual annotation. In-context learning consistently outperformed the other methods, suggesting possible label noise or limited representativeness in the synthetic data. Furthermore, clarity emerged as the most linguistically unpredictable style. This study provides an initial framework for evaluating language style in LLMs for customer service, with the broader goal of improving alignment with real human communication.

Contents

1 Introduction **3**
1.1 Paper Structure . . . 4
1.2 Ethical Statement . . . 5

2 Related Work **5**
2.1 What Is Language Style? . . . 5
2.2 Language Style in Customer Service . . . 7
2.2.1 Clarity . . . 7
2.2.2 Friendliness . . . 8
2.2.3 Empathy . . . 10
2.2.4 Politeness . . . 11
2.2.5 The Issue of Stylistic Overlap . . . 14
2.3 AI in Customer Service . . . 14

3 Experimental Setup **16**
3.1 The RESPONSible Service Dataset . . . 16
3.2 Linguistic Data Profiling . . . 18
3.3 Fine-tuning LLMs on Synthetic Data . . . 20

4 Results and Discussion **23**
4.1 Linguistic Data Profiling . . . 23
4.1.1 Frequency Analysis . . . 23
4.1.2 Lexical Analysis . . . 26
4.1.3 Discourse Analysis . . . 27
4.1.4 Analysis of Deixis and Modality . . . 28
4.1.5 Sentiment and Emotion Analysis . . . 30
4.1.6 Additional Metrics . . . 32
4.2 Fine-tuning LLMs on Synthetic Data . . . 33

5 Conclusion **35**
5.1 Main Findings . . . 35
5.2 Limitations . . . 36
5.3 Future Work . . . 37
5.4 Final Remarks . . . 37

Appendix **42**

A Example Rows from the RESPONSible Service Dataset **42**

B List of Cohesive Devices Used in the Discourse Analysis **43**

C Prompts for the Text Generation Tasks **44**
C.1 In-Context Learning (ICL) . . . 44
C.2 Supervised Fine-Tuning (SFT) . . . 44
C.3 Direct Preference Optimization (DPO) . . . 44
C.4 Baseline (Zero-shot Chat Template) . . . 44

1 Introduction

Language style is a fundamental aspect of human communication. While style is commonly defined as a distinctive manner of expression (Irvine, 2001; Merriam-Webster, n.d.-f; Oxford Learner's Dictionaries, n.d.), it is not a purely linguistic concept, but also has important implications in relation to social norms (Irvine, 2001). Language style contributes to shaping not only how individuals are perceived (Irvine, 2001), but also how they construct and express their identities (Kiesling & Schilling-Estes, 1998). The use of specific words, punctuation, syntactic structures, or pronunciation patterns can indicate an attempt to adapt to the social context, but can also be a way of defining oneself in relation to such context (Kiesling & Schilling-Estes, 1998). The effects of language styles are not limited to individual language users, but can also have implications for larger entities, such as companies or organizations. One such case is the area of customer service (CS). When companies interact with customers, the way in which a message is phrased can impact customer satisfaction, and the reputation of the entire company in turn. Wrong stylistic choices can lead to lost customers and bad reviews, whereas careful and strategic CS communication can positively reflect on revenue. Different stylistic choices will have different effects. Unclear and vague responses to CS inquiries may cause irritation in the customers (Lutzky, 2024), while responses containing concrete, precise and specific information will not only be appreciated, but also create a perception of empathy by making the customers feel that they are being listened to (Packard & Berger, 2021). When CS agents include linguistic elements such as greetings and small talk (Creelman, 2022; Zabava Ford, 1999), as well as inclusive pronouns (Creelman, 2022; Ruytenbeek & Decock, 2024) and verbs in the present tense (Creelman, 2022), this contributes to reducing the social distance between the customer and the CS agent, and to creating a friendly and intimate atmosphere, which in turn will make the interaction more pleasant. Language style must also be calibrated according to demographic and cultural factors, as the same style does not guarantee the same outcome with every customer. For example, older people might prefer a more formal and detached approach, and the use of informal greetings or exceedingly direct invitational expressions may be perceived negatively (Creelman, 2022). Similarly, the same style resonates in different ways in different cultures. The preference for positive politeness of American CS agents may be perceived as disingenuous by non-American customers. This might be the case with British people, who are more accustomed to negative politeness, thus making greater use of hedges and indirect phrasings (Cameron, 1997; Pinto, 2011). Furthermore, a single stylistic choice can have a range of different implications, which more often than not are hard to disentangle, especially in light of the aforementioned demographic and cultural differences.

The growing use of automation and artificial intelligence (AI) when communicating with customers has led to improved CS interactions, especially with the development of AI-based chatbots. These technologies are used not only to solve technical issues, but also to engage customers on a more personal level. Qualities such as empathy, politeness, and friendliness are often addressed when designing tools for customer interaction, as they have been found to significantly influence satisfaction and trust. However, these qualities are not implemented in the same way, with some of them receiving more attention than others. Furthermore, the overlap between many qualities makes it difficult to address them in isolation. For these reasons, although AI models have made progress in producing natural emotional responses, they often fail to deeply understand customer emotions or to provide context-appropriate reactions (Mendonça et al., 2023). Rather than completely replacing human CS agents, some approaches have attempted to support them by simplifying their work. For instance, AI models have been used to moderate emotionally charged language or to help staff cope with difficult interactions (Jo & Seo, 2024; Swain et al., 2024). These tools, while helpful, also show the challenges that AI faces in replicating the nuance of human emotional intelligence. Customer responses to AI are influenced by how well these systems align with human values. For example, empathetic language helps build customer trust, but when used inappropriately, it can reduce the perceived reliability of the system and cause frustration (Han et al., 2022). Or a friendly tone, while generally appreciated and linked to increased trust, may not be sufficient on its own when technical competence is required (Cheng et al., 2021). Furthermore, some qualities remain underresearched. Communication clarity is one of them, despite the fact that lack of clarity has important implications in CS success. Unclear responses may frustrate users and reduce service effectiveness, even if the tone is warm or understanding (Lutzky, 2024). All in all, there is still limited research on the intersection of language style, CS, and AI. One of the potential causes of this could be the lack of publicly available datasets that focus on language style in CS communication. One can hypothesize that private companies are reluctant to share their CS data and prefer to use them for internal purposes. At the same time, lack of suitable data for specific tasks has been increasingly addressed by resorting to synthetic data generation through large language models (LLMs). While this approach offers a practical workaround, it also raises concerns about how accurately synthetic data reflect the nuances of real human communication.

Therefore, in this study, I will examine how LLMs represent language style in CS by creating and analyzing a synthetic dataset of customer requests and stylistically marked CS responses. The dataset will focus on the styles of clarity, friendliness, empathy, and politeness. I will analyze the dataset by conducting a profiling of the linguistic aspects that characterize each of the aforementioned styles, and I will interpret the results in light of prior research on human CS communication. Additionally, I will use the synthesized dataset to train models for text generation in different fine-tuning settings (in-context learning, supervised fine-tuning, and direct preference optimization) and evaluate their outputs in terms of semantic and formal overlap with the ground truth using BERTScore and ROUGE-L, as well as in terms of human agreement through manual annotation. This will provide a partial assessment of the quality and usability of the synthetic dataset.

With this study, I want to answer the following research questions (RQs):

RQ1: What linguistic patterns emerge in LLM-generated customer service responses when the LLM is prompted to reflect the styles of clarity, friendliness, empathy, or politeness?

RQ2: How do these patterns differ across the four styles?

RQ3: In what ways do these patterns align with or diverge from those discussed in prior research on human customer service communication?

RQ4: How can the text generation performance of models adapted to LLM-generated customer service data be evaluated, and what insights emerge from this evaluation?

Given the limited existing research in this specific area, this study will take an exploratory approach. My goal is to shed light on how LLMs represent and distinguish CS-related language styles. I hope that my work will serve as a foundation for future research on how LLMs can be stylistically aligned with human CS communication, thereby improving customer satisfaction and trust.

1.1 Paper Structure

I will start by examining the literature regarding language style in general and in CS in particular, as well as the use of AI in CS, in Section 2. In Section 3, I will proceed to discuss the creation of the synthetic dataset, and the methodology behind the linguistic data profiling and the text generation tasks. In Section 4, I will present and discuss the results of the experiments. Finally, in Section 5 I will draw my conclusions and suggest further research.

1.2 Ethical Statement

This study did not involve human participants, personal data, or any procedures requiring ethical approval. All data used in this research were synthetically generated using GPT-4o, with no reliance on identifiable or sensitive human information. In accordance with best practices in language research, I also note that all generated data are in English.

2 Related Work

2.1 What Is Language Style?

Language is a core element of human experience, and for this reason language style has strong implications on a person's life. But what is style? According to the Merriam-Webster dictionary, style is defined as "a distinctive manner of expression (as in writing or speech)" (Merriam-Webster, n.d.-f). The Oxford Learner's Dictionary defines style in general terms as "the particular way in which something is done," but it also provides a more language-specific definition, describing style as "the correct use of language" (Oxford Learner's Dictionaries, n.d.). As we can see, language style is not just a specific way of speaking or writing, but also one that is generally considered correct. However, *correct* is also a generic term that refers to a relative concept. If we interpret correctness as social appropriateness (Merriam-Webster, n.d.-b), further clarification regarding the interrelatedness between language style and correctness is provided by Irvine (2001). Irvine defines style not just as a specific way of expressing language, but a full-fledged sociolinguistic phenomenon that has a strong connection with societal processes and whose purpose is to create social differentiation. Therefore, style is not a fixed construct, but rather a process that keeps evolving and adapting to the social context to which it applies. This means that different social situations will require different styles, and styles that may be considered appropriate in a certain situation, may not be in a different one. Irvine mentions the case of two language styles used by German speakers in Bóly, a small town in Hungary. Two styles are distinguished: the farmer style and the artisan style. The differences between the artisan and farmer language varieties surface throughout the entire linguistic system. These include variations in pronunciation, grammar, vocabulary, and broader patterns of discourse. The artisan style appears more innovative, as it tends to incorporate more external influences and borrowed elements from other dialects and languages. In contrast, the farmer style remains more conservative and local. As anticipated, the difference between the two styles is not merely linguistic, but also social. While the farmer style is associated with values such as frugality and sobriety, the artisan style embodies sophistication and elegance. German speakers from Bóly are proficient in both styles and use them interchangeably based on factors such as social context, interlocutor, and mood.

According to Kiesling and Schilling-Estes (1998), language style is not just a way for the speaker to adapt the language to a social context, but it also contributes to the construction and reconstruction of the speaker's identity. In fact, personal identity is not innate but actively constructed by means of social interaction and by emulating the behavior of others (Luckmann, 2008). Kiesling and Schilling-Estes maintain that speakers use style to put themselves or others into specific roles, or *footings*. These footings allow speakers to position themselves in relation to social actors or social groups. This process takes place in a specific speech activity, or *frame*, of which all the different parties involved in the interaction are usually aware. Footings do not necessarily refer to existing demographic groups, but also to social prototypes, such as the *physically powerful man* or the *quaint island fisherman*. The authors explain these concepts by means of a study that was conducted in a fraternity at a Virginia college in relation to the use of the alveolar nasal consonant as opposed to the velar one in English -ING forms, and in different social contexts, such as interviews, fraternity meetings, and casual social events. It was found that the participants of this study did not adapt their language simply in relation to a generic idea of the social context in which they were, but rather in relation to their own framing of such context, and in doing so they indexed specific social prototypes related to such framing. This intuition emerged from the fact that not all students used the same pronunciation of the -ING form in the same contexts. The students who framed the meetings as serious leadership spaces and wanted to be perceived as more authoritative and rational, thus looking up to prototypes of people who hold power (such as professors and CEOs), generally preferred the standard pronunciation. In contrast, the students who framed the meetings as collaborative spaces and attributed more value to footings based on solidarity and physical strength, for instance by identifying in the prototypes of the *hard-working and working-class men*, used the non-standard alveolar pronunciation more often. This examples shows us that style choice is an active choice rather than a passive one, and it is strongly connected not only with the way other people perceive us, but also with the way we perceive our own identity.

The use of style to define one's identity can be applied not only to individuals, but also to larger social entities such as companies and organizations. For example, in the corporate world, corporate voice and conversational human voice (CHV) (Kelleher, 2009) are two commonly found language styles (Barcelos et al., 2018). The former is the traditional one, and is usually formal and detached, whereas the latter is more innovative and uses a friendly and informal style. CHV is defined by Kelleher (2009) as "an engaging and natural style of organizational communication as perceived by an organization's public based on interactions between individuals in the organization and individuals in the public". The choice between the two tones of voice depends on the context in which corporate communication happens. According to Barcelos et al. (2018), CHV is suitable when the consumer's goal is hedonic and the situational involvement is low, such as in the case of a vacation trip. In contrast, corporate voice is preferable when the consumer goal is utilitarian and the situational involvement is high, which could be the case of a business trip. Tone choice can also be interpreted as a way for a company or an organization to actively define its role in society in general and in relation to the consumer more specifically. The Italian company Fiscozen, which provides tax accounting services for freelancers, adopted a conversational, straightforward, and empathetic language style that is not commonly associated with taxes and more in general with any government-related matter (Fiscozen, n.d.). Tax accounting is traditionally characterized by a highly technical and specialized language (McIsaac & Sepe, 1996), which can be difficult to understand for laymen. It has been proven that this type of language can be perceived as more authoritative and captivating (Weisberg et al., 2008). However, the choice of such style can also backfire. According to Oppenheimer (2006), the author of a text written using complex sentence structures and long words is considered less intelligent compared to the author of a text that uses clear language and syntax. Furthermore, Reber et al. (2004) maintain that stimuli that can be processed more easily are evaluated more positively. When Fiscozen decided to revolutionize tax-related communication by choosing its current language style, it differentiated itself from the rest of the tax accounting world and took on the *footing* of the pioneer, indexing prototypes such as the innovator and the nonconformist.

2.2 Language Style in Customer Service

A specific area in which a company can significantly express its identity is CS. In this area, a company constantly interacts with the public, and the chosen language style has a strong effect on the success of the exchange between the customer and the CS representative. The active choice of a carefully designed language style for CS can be a way for the company to take the lead in an important process that will further shape the way the company is perceived from the outside. Edvardsson et al., as cited in Lewis and Mitchell (1990), maintain that style, as a component of functional quality, is one of the aspects of CS that influence customer perceptions. In the following five sections, I will discuss language style in relation to four stylistic features (clarity, friendliness, empathy, and politeness), as well as the overlap between them. The choice of these four qualities is not arbitrary. They were chosen because they are representative of most of the qualities that according to the literature are considered essential for successful CS communication. In the following sections, I will discuss each quality individually and explain how it relates to previous findings.

2.2.1 Clarity

While clarity is important in all forms of communication, it is especially crucial in CS. Clarity can be conveyed through elements such as word choice, sentence structure, and reference specificity. Zabava Ford (1999) discusses the role of language clarity in relation to the concept of *personalized service*, which is one of the three types of CS communication style that she identifies, the others being *courteous service* and *manipulative service*. One key aspect related to clarity is the practice of information sharing. Zabava Ford explains that personalized service means conveying technical or unfamiliar information in a way that customers can understand. For example, she observes that physicians frequently use technical terms believing their patients will understand them. However, this is usually not the case and often leads to confusion. Effective information sharing requires using simple language, framing techniques, and multilevel explanations. The strategy of framing refers to the act of paraphrasing information in ways that can be more easily understood by the interlocutor, whereas multilevel explanation refers to the act of expressing complex concepts followed by summaries that use a simpler language. These behaviors are strongly connected with the concept of communicative clarity, as they are meant to enhance customers' understanding. Another concept tied to clarity and personalized service is *interaction involvement*, which includes attentiveness, perceptiveness, and responsiveness. Attentive and perceptive providers can better understand what the customer is trying to say, enabling them to respond in ways that are relevant and easy to follow (responsiveness), thus improving clarity.

Conversely, clarity can be compromised when CS agents resort to *manipulative service* strategies, which Zabava Ford associates with control and deception. For example, providers may intentionally avoid certain topics or control the flow of interaction through scripted questions and prompts rather than addressing the customer's actual concerns. This is done for various business-oriented reasons, such as increasing sales or speeding up service, and might result in vague or irrelevant communication, thus negatively affecting clarity. This aligns with the findings of Lutzky (2024), who conducted a study on a one million-word corpus of Ryanair Twitter interactions, showing that a lack of clear, specific information is one of the main causes of customer dissatisfaction. Many customers complain about the absence of clear communication, explanations, and updates, especially in situations like flight delays or booking issues. Collocations with the negative particle "no" (e.g., "no explanation", "no info") reveal that passengers frequently feel left in the dark. Even after receiving corporate replies, customers are often forced to repeat their questions, indicating that the initial answers were unclear, irrelevant, or too generic. For instance, customers criticize "standard replies" that do not address their specific issues, and sometimes they even question whether the replies are automated. Furthermore, they express frustration for not receiving a "definitive" or "straight" answer, showing a clear preference for direct and relevant communication. Customers appreciate "quick" replies, but speed alone is insufficient if the content is vague or generic. Lutzky also finds that formulaic language, like "please contact us DM in order to assist you better", is frequently used by Ryanair and is perceived negatively when it does not solve the customer's problem. Lutzky maintains that organizations should move away from template responses, and adopt context-sensitive and informative communication strategies to enhance customer satisfaction and improve brand perception.

Moving back to Zabava Ford, another CS communication style she discusses is *courteous service*, with one of its main components being *verbal immediacy*. Language can indeed be used to create a sense of psychological closeness between customers and CS representatives. One example of this is the strategic use of spatial expressions. In particular, proximal adverbs and demonstratives such as *here*, *this*, and *these* convey more immediacy compared to distal ones such as *there*, *that*, and *those*. I maintain that in certain cases verbal immediacy not only contributes to perceived closeness, but also to communicative clarity. For example, Piwek et al. (2008) found that, in Dutch, proximal deictics are more accessible than distal ones, with accessibility defined as "the ease (of effort) with which particular mental contents come to mind." The relation between clarity, reference, and customer satisfaction has also been discussed by Packard and Berger (2021), who conducted different studies in which they demonstrated that using concrete language when communicating with customers increases satisfaction and purchase rate. Concrete language refers to language that is specific, tangible, and clear, such as, for example, asking a customer whether they would like a *coffee* as opposed to whether they would like *anything*. The authors found that the use of concrete language is perceived by customers as an indicator of the worker's ability to listen and to modify the service according to the customer's needs. This shows how concrete communication is closely related to language clarity as well as to other important elements of CS communication, such as attentiveness and responsiveness.

2.2.2 Friendliness

In CS communication, friendliness is commonly seen as a key element of positive interactions. In their respective studies, Zabava Ford (1999), Creelman (2022), and Ruytenbeek and Decock (2024) share the idea that friendliness is not just a personal trait, but also an intentional linguistic performance that influences how customers perceive both the service provider and the company as a whole. Zabava Ford (1999) defines friendliness as a fundamental feature of courteous service. Courteous service is also explicitly defined as *friendly service*, and is characterized by specific verbal and nonverbal cues that help create rapport and reduce psychological distance. Zabava Ford breaks down the verbal cues into two main categories: phatic speech and verbal immediacy. Phatic speech, such as greetings, expressions of gratitude (e.g., "Thank you"), and small talk, has the function of creating social bonding. These cues are particularly powerful because of their positioning at the beginning and end of interactions, thus leveraging the primacy and recency effects, psychological principles that suggest people remember the first and last elements of an experience more vividly. For this reason, Zabava Ford emphasizes that greetings and closings play an important role in shaping the customer's perception of the service. Verbal immediacy, instead, consists of linguistic strategies that indicate presence, engagement, and relational closeness. These include the use of present-tense verbs (e.g., "I'm working on your issue" instead of "I worked on your issue"), inclusive personal pronouns (e.g., "Let's see what we can do") as opposed to exclusive ones (e.g., "You need to do this"), and proximal spatial deictics as opposed to distal ones (e.g., "here" and "this" vs. "there" and "that"), all of which contribute to a perception of immediacy and shared experience. Such subtle choices help create a more collaborative tone, making interactions less transactional or hierarchical. Zabava Ford also highlights other forms of verbal friendliness, such as using the customer's name, acknowledging their concerns directly, expressing empathy, and even using humor, though she also notes that excessive friendliness, particularly in the form of prolonged small talk, can conflict with customer expectations for efficiency, particularly in time-sensitive contexts.

Creelman (2022) similarly emphasizes the role of language in expressing friendliness, focusing on digital CS interactions. She identifies a four-move pattern that is commonly found in corporate responses: greeting, expressing appreciation, delivering information, and closing. Each of these moves provides opportunities to integrate friendliness through lexical and stylistic choices. Greetings such as "hi" or "hey" paired with the customer's first name set an informal and personalized tone right from the start of the interaction. Appreciation moves, which often include phrases like "Thanks for reaching out," show goodwill and validate the customer's position, thus helping to portray the company as attentive and responsive. Creelman also observes that expressions featuring positive intensifiers, interjections, and exclamation marks, such as "That's great to hear!" or "That's exciting!," can be used to show involvement and even agreement with the customer, thus reinforcing the common ground. The use of emoji (especially smiley faces) further contributes to the impression of friendliness, functioning as a digital proxy for nonverbal cues like facial expressions. Creelman's analysis is particularly useful in showing how friendliness can emerge in written digital contexts, where traditional nonverbal cues are absent. The incorporation of paralinguistic elements such as punctuation, emoji, and expressive lexical items creates a tone that mimics spoken conversation and helps to humanize the interaction. Her findings show that, while these cues are often informal, the tone they help create is not necessarily casual or unprofessional. Instead, they can be adjusted to convey a semi-formal warmth that suits the brand's identity, as in banking (Creelman, 2022) or taxation (see Section 2.1), where friendliness is balanced with institutional credibility.

Ruytenbeek and Decock (2024) complement Zabava Ford's and Creelman's studies by addressing friendliness as a component of CHV (see Section 2.1) and as a culturally-specific factor. According to Ruytenbeek and Decock, CHV can be characterized in terms of three strategies: *message personalization*, *informal register*, and *invitational rhetoric*. Message personalization involves using first- and second-person pronouns (e.g., "I," "you," "we") and addressing the customer by name, strategies that signal direct engagement and reduce social distance. Informal register encompasses the use of casual vocabulary, interjections, emojis, and sometimes colloquial expressions, which help emulate the relaxed tone of face-to-face interaction. Invitational rhetoric refers to the practice of explicitly manifesting one's intention to initiate an interaction, for example by asking the customers for feedback. This approach positions the company as open, attentive, and eager to engage. Ruytenbeek and Decock also highlight the cultural variability in how friendliness is expressed. Their cross-linguistic analysis shows that English-language CS responses are typically richer in interpersonal features like gratitude and empathy, while Spanish responses are more task-focused and impersonal. This insight reveals that linguistic markers of friendliness are not universal, but are shaped by social and cultural expectations about service, professionalism, and emotional expression. All three studies converge on the understanding that friendliness is not simply about being nice. It is a deliberate, often strategic, deployment of linguistic and stylistic features to create emotional resonance, reduce social distance, and foster trust. These cues function across various levels, such as the lexical level (e.g., word choice), the morphosyntactic level (e.g., use of tense and pronouns), the paralinguistic level (e.g., punctuation and emoji), and the pragmatic level (e.g., responsiveness and topical alignment). When used effectively, they contribute to what customers perceive as authentic, human-centered interaction, even in asynchronous formats.

2.2.3 Empathy

While empathy in CS may surface through explicit emotional expressions such as apologies, most often it appears to emerge subtly through more implicit linguistic choices. Such linguistic cues seem to significantly affect how well customers feel heard, understood, and valued. In Section 2.2.1, I introduced the findings of Packard and Berger (2021), which suggest that the use of concrete language, as characterized by specific, tangible, and vivid words, plays a positive role on customer satisfaction. However, the authors also found that this seems to be the case only when such concreteness signals that the employee is actively listening, that is, attending to and understanding the customer's needs. In a controlled experiment, participants imagined contacting customer service to add gray pants to an order, and then read a scripted employee reply that varied by (1) language concreteness (concrete vs. abstract) and (2) response relevance (accurate vs. inaccurate). The results were clear: concrete language increased satisfaction only when the employee's reply was relevant. When the employee misunderstood the request, concreteness no longer helped and even negatively affected perceptions. A moderated mediation analysis confirmed that this effect was explained by changes in perceived listening. That is, concrete language led to higher satisfaction only when it made the customer feel heard. The authors tested alternative explanations, such as perceived caring, perspective taking, and empathetic concern, and found that these did not directly account for the effect of concrete language on satisfaction. However, they noted that these factors do help explain the impact of perceived listening on customer satisfaction. Indeed, the perception of empathy may emerge as a consequence of feeling listened to. When employees use relevant, concrete language that signals understanding, customers may interpret this as empathetic behavior.

Clark et al. (2013), through their analysis of call-center interactions, articulate the pragmatic nature of empathy that emerged in Packard and Berger (2021)'s study through the notion of *empathy work*, categorizing empathetic communication into three distinct types: *attentive*, *affective*, and *cognitive*. Attentive empathy involves direct and minimal signals of active listening, such as brief acknowledgments ("Yes, I understand, ma'am"), clarification requests ("Do you actually have any agent in mind?"), or concise summarizing statements ("Okay, so what I'll do is..."). In contrast, affective empathy includes general emotional expressions such as apologies or standard phrases of sympathy ("That must be frustrating"). Finally, cognitive empathy is characterized by deliberate attempts to understand and assist customers in fulfilling their needs. For example, when a customer struggled to recall a product name ("the medical program, the gold, ah..."), the agent immediately provided precise assistance ("The Golden Years Plan, aha"), showing a genuine understanding of the customer's situation. Clark et al. emphasize that, while affective empathy often came across as superficial or insincere, cognitive empathy was the most positively impactful. By offering targeted solutions or clarifying confusing terminology, cognitive empathy effectively reassures customers that their needs have been understood. These findings closely align with the ones by Tan et al. (2019) in the context of higher education. The authors examined empathy by surveying 256 students and staff from 11 Singaporean universities. They found that empathy is important to both students and staff, particularly the provision of individualized attention to students, which was found to be the strongest positive predictor of student satisfaction. However, the idea of empathy of students and staff diverge in significant ways. Students prioritize whether staff have their best interests at heart, while staff emphasize their knowledge of students' needs, and their ability to show care and concern. Interestingly, students react negatively to excessive expressions of care, which they may perceive as "over-servicing" or "spoon-feeding", a practice that can reduce student autonomy and engagement. Students seem to prefer professional, respectful, and individualized communication over overtly friendly or emotionally loaded interactions, which may be seen as inauthentic or patronizing. Tan et al.'s findings demonstrate that the best communicative strategy when it comes to empathy is highly dependent on the context and on the expectations of the interlocutor. At the same time, they also demonstrate that emotional language is not necessarily the correct approach to convey empathy.

Ngo et al. (2020) further enrich this understanding by highlighting empathy's reciprocal nature through practical examples in retail service encounters. The authors surveyed 211 pairs of front-line employees and customers in Vietnamese securities companies. Their analysis demonstrated that empathy only leads to higher customer satisfaction when customers themselves are emotionally and cognitively invested. This investment may be manifested in various way, such as by showing appreciation, articulating clear needs, and offering constructive feedback. Without such reciprocal interaction, the benefits of employee empathy diminished. This shows that empathetic service delivery is not unidirectional, but cooperatively constructed in the moment, relying on mutual signals of attentiveness and care. In a similar way, Wieseke et al. (2012) offer a multilevel analysis of how empathy functions between employees and customers, introducing the notion of *symbiosis* to describe mutual empathetic alignment. Their study, based on dyadic data from travel agencies, found that the impact of employee empathy on customer satisfaction was significantly enhanced when customers also exhibited empathy. Moreover, the study also showed that empathetic customers were more likely to forgive service failures, particularly when they perceived that employees were making a sincere effort. In this way, empathy emerges as a psychological mechanism that not only improves satisfaction, but also softens the negative effects of dissatisfaction on customer loyalty. In summary, across these five studies, the common insight is that empathy in CS communication is less about overt emotionality and more about context-sensitive linguistic practices. Empathy is also not a fixed behavior, but a set of subtle interactive adjustments that, when executed skillfully, create the sense of a service that is genuine and focused on the person rather than on the business.

2.2.4 Politeness

Politeness plays a central role in CS communication. Effective use of politeness strategies can reduce conflict, and improve satisfaction and customer loyalty, whereas ineffective expressions of politeness may lead to misunderstandings or dissatisfaction. But what actually is politeness? When we think about politeness, the first thing that comes to mind are words and phrases such as "please" and "thank you." However, politeness extends far beyond everyday routine expressions. It encompasses a wide range of strategies used to manage face and maintain social harmony in interaction. The framework introduced by Brown and Levinson (1987) distinguishes between positive politeness, which builds rapport and solidarity, and negative politeness, which minimizes imposition. These concepts represent the foundation of most current research on the role and expression of politeness in communication. As we saw previously in the study by Creelman (2022), CHV in CS is achieved not merely through informality, but through a structured use of language that balances professionalism with warmth. Politeness has a central role in this balance. Drawing on Brown and Levinson's (1987) framework, Creelman shows how positive and negative politeness strategies help manage the relational dynamics of online service interactions, especially in the absence of nonverbal cues. Due to the nature of positive politeness, which is based on rapport building, it is not always easy to distinguish it from friendliness. Many of the same linguistic devices used in positive politeness to validate the customer's message and create a sense of mutual alignment, such as positive exclamatory statements (e.g., "Thank you for reaching out!") and emoji, are also found in friendly language (see Section 2.2.2). In contrast, negative politeness refers to the customer's desire for autonomy and to not be imposed upon. This is evident in conditional and indirect phrasings like "PM us if you'd like to schedule an appointment." Such language softens imposition and commands, and thus supports a tone that is helpful yet respectful. Importantly, Creelman observes that these politeness strategies are not generic. They are calibrated to the context, such as the seriousness of a complaint or the preferences of different customers, which can depend on various factors, especially age. Friendliness and politeness, negative politeness in particular, both function as core components of CHV, allowing agents to make online interactions feel more human while maintaining a tone that is respectful and consistent with the context, thus fostering credibility and professionalism.

However, what politeness amounts to exactly is often difficult to define with precision, and cultural variability has a central role in this. Cultural norms strongly influence how politeness is expressed and interpreted. Different cultures attach varying degrees of significance to indirectness, honorifics, and formulaic expressions, which can lead to mismatches in perception during intercultural interactions. Pinto (2011) explored American politeness norms and highlighted the widespread use of formulaic politeness markers such as "please," "thank you," and "I'm very sorry," which are deeply embedded in everyday American discourse. These expressions, although they generally aim to convey friendliness, warmth, and social harmony, can nonetheless be perceived by outsiders as superficial or insincere. Pinto points out that the abundance of these routinized expressions may convey an exaggerated sense of emotional involvement that does not always align with the speaker's true intentions or feelings. For instance, a simple greeting like "Hi, how are you?" at a grocery store checkout, though delivered politely, often does not solicit or expect a genuine answer. As some participants in Pinto's study noted, this phrase has more the function of a social signal to start interaction than as a sincere interest in the customer's well-being. Similarly, questions like "Did you find everything you were looking for?" are often asked at a point in the transaction (during checkout) when no meaningful assistance can be offered, which reinforces the perception that such questions are performative rather than genuine. Pinto draws attention to cultural clashes that can arise when Americans interact with individuals from cultures where politeness is expressed differently, often with more restraint or through actions rather than words. For example, the high frequency of compliments in American speech, particularly regarding personal appearance, or the American tendency to indirect refusals ("Oh, thank you ever so much for the invitation but I have to go and see my parents") can strike Japanese and German speakers as shallow, due to differing expectations around directness and sincerity. Overall, Pinto suggests that American politeness often revolves around *rapport-based* sincerity, where the primary goal is to maintain positive social relations and make interactions smooth and pleasant, even at the cost of literal honesty. This is in contrast to *truth-based* sincerity, more common in some other cultures, where the authenticity of one's words is essential. As a result, while Americans may view their polite routines as courteous and well-meaning, interlocutors from other cultural backgrounds may perceive the same behavior as insincere or forced.

Friginal (2024) offers an additional perspective on the cultural aspects of politeness. The author investigated the variation in politeness in call center CS agents from three different countries: US, India, and Philippines. Using corpus-based analysis, the study shows that Filipino agents consistently use politeness markers such as "please," "thank you," "sir/ma'am," and apologetic expressions like "I apologize for the inconvenience," more frequently than their Indian and U.S.-based counterparts. These features co-occur with longer and more elaborated turns as well as a larger use of second-person pronouns, indicating a customer-centered and respectful approach to service. In contrast, U.S.-based agents tend to be more direct, using shorter transactional responses with fewer politeness expressions. Indian agents show a hybrid pattern, between the more direct U.S. style and the more tactful Filipino style. An example from the paper highlights these differences: to ask the customer for their phone number, the Filipino agent says: “Please give me your cell phone number so I can check on your minutes.” This contrasts with the curt approach of the U.S. agent, who simply says: “I just need your number” without any additional elaboration. These findings suggest that cultural norms and local training practices influence how politeness is expressed in CS interactions, impacting customer satisfaction and perceived professionalism in call center communication. In this context, politeness can also often become standardized through scripts, which may be introduced to optimize CS interaction, but often end up paradoxically hindering it. Cameron (1997) argues that the use of standardized expressions can significantly limit the flexibility of CS agents, who are forced to insert phrases that are not culturally neutral into the wrong context. Prescribed politeness routines can create a disconnect between form and intent, leading customers to doubt the sincerity of the interaction. Back in 1997, Cameron observed how American CS language was already influencing the way in which CS agents in the UK interacted with customers. This linguistic influence was not limited to the replacement of typically British expressions with more American ones, such as in the case of “How may I help you?” (US) replacing “Can I help you?” (UK), but also extended to the export of a model of politeness that was incompatible with British society. Indeed, according to Cameron, American speakers are more inclined to use expressions signaling positive politeness compared to British speakers, who favor expressions of negative politeness. A mismatch between the type of politeness that is used and the context in which it is used may cause unpleasant consequences, such as politeness expressions being perceived as demeaning rather than friendly.

Another area of development for CS communication is social networks, whose advent has expanded the repertoire of ways in which politeness is expressed and perceived. On platforms like Twitter and Facebook, public visibility and efficiency shape both customer complaints and corporate responses. Lutzky’s (2024) study on Ryanair’s CS interactions, which was also discussed in Section 2.2.1, provides further insight into the role of politeness in social media, especially in its intersection with clarity. The airline’s replies, which frequently followed standardized templates without necessarily providing the required information, were often perceived by customers as generic, delayed, or unhelpful. Fuoli et al. (2021), as cited by Lutzky (2024), observes that customers might see generic responses as impolite because they can come across as distant or impersonal, giving the impression that the company does not really care about their issue.

As previously observed in the case of empathy, politeness also emerges as a reciprocal process. Hu et al. (2019) found that customers using polite linguistic cues, such as indirect requests (e.g., “By the way, I also want to...”) or hedges (e.g., “I’m not an expert, but...”), were more likely to receive public replies from companies. In contrast, messages with aggressive or impolite language were often moved into private channels, demonstrating how politeness influences response strategy and visibility.

In conclusion, politeness in CS cannot be reduced to a set of fixed expressions, but must instead be flexibly adapted to interactional, cultural, and contextual variables. The studies reviewed in this section demonstrate how politeness markers serve not only to manage face, but also to shape perceptions of authenticity, empathy, and professionalism in an increasingly complex service landscape. As digital communication continues to evolve, understanding the linguistic nuances of politeness will remain central to delivering effective CS.

2.2.5 The Issue of Stylistic Overlap

As we have seen, clarity, friendliness, empathy, and politeness are essential for improving customer experiences in CS interactions. While each of these styles has distinct features, they often overlap in subtle and complex ways, making them difficult to disentangle in practice. This interconnectedness is evident in both theoretical discussions and empirical studies, which consistently show that these communicative behaviors often co-occur, support, and reinforce each other within the flow of service interactions. At its heart, clarity means conveying information effectively, and it is achieved through simple, concrete, and contextually relevant language (Lutzky, 2024; Packard & Berger, 2021; Zabava Ford, 1999). However, these elements can also affect perceived empathy by conveying attentiveness and ability to listen (Clark et al., 2013; Packard & Berger, 2021; Zabava Ford, 1999), as well as politeness, since irrelevant and formulaic responses may be perceived as impolite (Fuoli et al., 2021; Lutzky, 2024). Creelman (2022) and Ruytenbeek and Decock (2024) show that friendliness in online CS interactions is often built into the structure of corporate communication through personalized greetings, appreciation moves, and expressive punctuation. As we saw previously, these strategies, which are designed to convey warmth and attentiveness, can also be markers of positive politeness (Brown & Levinson, 1987; Pinto, 2011). Negative politeness, instead, could be easily associated with empathetic behavior as, by making use of hedges and indirectness, it shows consideration for the interlocutor's autonomy and emotional state. Creelman (2022) further illustrates how politeness, friendliness, and empathy overlap in digital contexts, where phrases like "Thanks so much for reaching out!" simultaneously express gratitude (a politeness strategy), emotional enthusiasm (a friendliness strategy), and interpersonal engagement (a signal of empathy). When cross-cultural factors are also taken into account, isolating the dimensions of clarity, friendliness, empathy, and politeness becomes even more complex. For example, while friendliness and politeness may overlap in certain cultures (Friginal, 2024; Pinto, 2011), they may be perceived as more distinct in others (Cameron, 1997; Pinto, 2011). In conclusion, clarity, friendliness, empathy, and politeness in CS are best understood not as discrete categories but as overlapping ones, which together shape the customer experience. While each has distinct theoretical foundations, in actual communication they often emerge linguistically and behaviorally in very similar ways. Greeting someone by name, using concrete and direct language, acknowledging frustration, and expressing gratitude can all happen in the same sentence, and each act can simultaneously fulfill multiple communicative goals. This functional overlap can make it difficult to isolate which element is driving the interaction, but it also points to the richness and flexibility of human communication. As CS increasingly moves into digital, cross-cultural, and automated contexts, understanding the dynamic interplay among these styles is key to designing interactions that feel not only efficient, but also human.

2.3 AI in Customer Service

Companies are increasingly using automation and AI to make CS communication faster, easier to scale, and more consistent. Thanks to LLMs, AI can now hold more natural conversations and is expected to not just solve technical problems, but also connect with customers on a personal level. More and more researchers are looking into how AI can show key human traits like empathy, politeness, and friendliness, which all shape how customers feel about their service experience. However, not all of these traits get the same attention in AI systems, and one of the main components of communication, namely clarity, is still often overlooked.

Huang and Rust (2024) propose a four-stage *customer care journey*: *emotion recognition*, *emotion* *understanding*, *emotion management*, and *emotional connection*. This purpose of this framework is to evaluate AI's role in emotionally charged customer interactions. While they find that generative AI (GenAI) can identify customer emotions and generate responses that appear empathetic, they also note that such systems may struggle with genuine emotional understanding, sometimes generating responses that are contextually inappropriate or lack personalization. These challenges highlight the difficulty of designing AI that goes beyond surface-level cues and truly aligns with the customer's emotional needs. Mendonça et al. (2023) approach these questions through empirical analysis of real-world bilingual CS conversations, where the CS agent only speaks English and the customer only speaks either Portuguese or German. Their MAIA dataset includes detailed annotations for dialogue quality and emotional features, including empathy and politeness. They report that state-of-the-art models, while effective at recognizing more common emotional cues such as neutrality or basic empathy, often perform more poorly when identifying more subtle or negative emotions like frustration or disappointment. In terms of politeness, their benchmarks show that current models struggle to predict this feature accurately, especially in dialogues where negative emotional dynamics are present. Their work emphasizes the need for better emotionally aware dialogue systems and reveals the challenges of maintaining conversational quality across diverse languages and emotional states.

Jo and Seo (2024) address the emotional stress that CS work can place on human agents. They propose a system, ProxyLLM, that uses an LLM to transform emotionally charged or aggressive customer messages into more neutral or polite forms before they reach the agent. ProxyLLM performs an empathetic function by protecting agents from distressing language and supporting their emotional well-being. Evaluations with multiple models show a clear shift toward a more positive tone. ProxyLLM also enables customizable style-transfer prompts, helping align customer communication with organizational standards. Jo and Seo's study highlights how LLMs are capable of exhibiting empathetic behavior, and how this can be used to foster emotionally intelligent workplace practices. In the same vein, Swain et al. (2024) explore how AI can support the emotional regulation of CS agents by engaging with them through empathetic messages. To do so, they developed Care-Pilot, an LLM-powered assistant designed to help workers cope with client incivility. In a comparative evaluation, AI-generated support messages were perceived by CS agents as more sincere, actionable, and emotionally helpful than those written by human coworkers. These messages contributed to the emotional regulation of CS workers by helping them reframe and recenter negative thoughts, and humanize customer behavior. However, participants noted that Care-Pilot's lack of lived experiences limited its ability to understand the full nuance of emotionally complex situations.

Cheng et al. (2021), instead, investigate how consumers respond to chatbot attributes such as empathy and friendliness in e-commerce environments. Their results show that both traits significantly enhance users' trust in the chatbot. Empathy, defined as the chatbot's ability to understand and respond to consumer needs, was found to have a particularly strong influence on trust. Friendliness, conveyed through positive tone and attitudes, also contributed to trust, though to a lesser extent. In particular, the study found that as task complexity increased, the positive effect of friendliness on trust decreased, suggesting that a friendly tone may not be sufficient when users expect technical competence in more demanding interactions. In contrast, the relationship between empathy and trust remained stable regardless of task complexity. Han et al. (2022) add even more depth to the discussion about AI-generated empathy. Their experimental study investigates how customers perceive chatbot empathy in two contexts: when the chatbot responds to a customer's negative experience, and when it apologizes for its own shortcomings. Results indicate that while the former improves customer evaluations, primarily by enhancing perceptions of warmth, the latter can actually reduce satisfaction by undermining perceived competence. These findings reveal that empathy in AI, unless it is targeted and context-sensitive, can be ineffective and even counterproductive. When a chatbot expresses empathy inappropriately, such as after its own failure, it may appear inauthentic and could further undermine the chatbot's perceived competence (Han et al., 2022).

Across the literature, a recurring theme emerges: while AI systems are improving at mimicking human emotions, generating responses that feel appropriate across diverse social contexts remains a significant challenge. Although increasing attention is being paid to interpersonal traits such as empathy, politeness, and friendliness, one crucial aspect of communication is often overlooked: clarity. None of the studies reviewed here examine the clarity of AI-generated responses, that is, how easy they are to understand, how logically coherent they are, and how free from ambiguity. This gap needs to be addressed, given that clarity is widely recognized as a core component of effective customer service (Lutzky, 2024; Packard & Berger, 2021; Zabava Ford, 1999). Equally important is the availability of high-quality training data that reflect these human-centered traits. Companies often collect CS data to monitor performance, identify trends, and refine their services. In some cases, these data are made publicly available for research purposes (Mendonça et al., 2023). In others, when suitable datasets do not exist, researchers create them by employing human annotators or writers (Gung et al., 2023). While effective, this process is time-consuming and costly (Busker et al., 2025). Recent advances in LLMs have opened the door to a promising alternative: the generation of high-quality synthetic data. These models, trained on massive corpora of human-authored text, are now capable of producing responses that are natural and contextually appropriate. As a result, researchers have begun exploring how LLMs can be used to generate training and evaluation data for smaller models (Busker et al., 2025; Sufi, 2024), in a process known as *knowledge distillation* (Braga et al., 2024). This synthetic data can either be derived by modifying real interactions (Shorten et al., 2021) or fully generated through carefully crafted prompts (Sufi, 2024). This approach allows for the use of the broad general knowledge embedded in LLMs while significantly reducing the costs and time associated with data collection. Still, synthetic data is only valuable if it meets high standards of quality. To ensure this, researchers typically rely not only on automated evaluation metrics such as perplexity, BLEU, and BERTScore (Gung et al., 2023; Zhang et al., 2019), but also on human annotators to assess the outputs (Sufi, 2024). When implemented rigorously, these methods can yield training data that enables AI to communicate in a manner that more closely resembles authentic human interaction.

3 Experimental Setup

The preceding sections examined the roles of clarity, friendliness, empathy, and politeness in shaping CS communication in both human and AI interactions. The following section introduces a synthetic dataset designed to isolate these communicative styles, with the aim of investigating how they are interpreted and reproduced by LLMs (Section 3.1). I then provide a methodological overview of the linguistic profiling conducted on the synthetic dataset (Section 3.2), as well as the use of this latter in the adaptation of a smaller language model to generate stylistically appropriate responses (Section 3.3). The implementation is available on GitHub.[1]

3.1 The RESPONSible Service Dataset

The RESPONSible Service dataset was created using GPT-4o via the AzureOpenAI API. The LLM was instructed four times to create data focusing on a different CS communication style each time:

[1]Available at: https://github.com/nicodebig/MA_Thesis_s2046059.git

clarity, friendliness, empathy, and politeness. These were chosen because they are representative of most of the qualities that according to the literature are considered essential for successful CS (see Section 2.2). Each style was assigned to a different subset in order to evaluate them independently of each other. For each subset, GPT-4o was prompted through a predefined response format, which was used to generate outputs consisting of three features each: `request`, `response0`, and `response1`. The following example, which focuses on clarity, shows how each feature was defined:

request: Request for assistance from a customer to customer service about {`random_topic`}. Generate exactly 30 words.

response0: Response from customer service to the customer's request. response0 must be phrased in a clear way. Generate exactly 30 words.

response1: Response from customer service to the customer's request. response1 must be phrased in a more clear way than response0. Generate exactly 30 words.

In earlier attempts, the model was required to produce outputs that either featured a certain quality or did not. This resulted in the model outputting highly polarized data, in which the two variants contrasted with each other in extreme ways. This was aggravated rather than improved by the presence of stylistic guidelines. Indeed, the presence of such guidelines seemed to cause the model to focus only on one idea of the respective style rather than on different possible nuances. For instance, the definition of clarity that was initially provided mentioned the avoidance of technical words. This resulted in the generation of *technobabble* when the model was required to be *not clear*. Defining each style in advance could have also introduced human bias in the model, by forcing it to adhere to the chosen interpretation of each style based on a limited amount of sources rather than to its underlying representations. For all these reasons, the prompt was revised to generate similar response pairs, with each item ranked based on its degree of adherence to the target style, rather than simply contrasting responses that did or did not exhibit stylistic cues. Furthermore, no style definition was provided, and the model was left free to attach its own interpretation to each style.

Each subset consisted of 1,000 items, with a total of 4,000 items across all four subsets. GPT-4o was required to adhere to a total output length of approximately 90 words for each generated example (30 words for each feature). The output length was kept quite low to contain computational costs and GPU usage (Chen et al., 2023). The `temperature` and `top_p` values were both kept at 1.0 for less predictable, more creative and more diverse outputs (Wang et al., 2023). Furthermore, to improve output diversity, the model was also forced to randomly select the topic of each generated example by cycling through a list of 11 different common CS topics, such as refunds, shipping, and discounts (hence the `random_topic` variable).

The dataset was also manually annotated to calculate human agreement and checked for potential leakage of sensitive data. However, due to time and resource constraints, this was carried out solely by the author. The manual annotation was carried out in the following way: for each item, the order of `response0` and `response1` was randomized, and a binary label was assigned: 0 if the responses were swapped, and 1 if not. The ground truth labels were then hidden, and the author independently labeled each pair by selecting the response that appeared more stylistically marked (0 for the response on the left, 1 for the response on the right). Human labels were subsequently compared to the original labels to calculate the human agreement ratio. The annotation focused on the following dictionary-based definitions of the four styles:

Clear: "Free from obscurity or ambiguity, easily understood." (Merriam-Webster, n.d.-a)

Friendly: “Of, relating to, or befitting a friend: such as a) showing kindly interest and goodwill; b) not hostile.” (Merriam-Webster, n.d.-d)

Empathetic: “Understanding, [...] aware of, [...] sensitive to, and vicariously experiencing the feelings, thoughts, and experience of another.” (Merriam-Webster, n.d.-c)

Polite: “a) Showing or characterized by correct social usage; b) Marked by an appearance of consideration, tact, deference, or courtesy.” (Merriam-Webster, n.d.-e)

The subsets for friendliness, empathy, and politeness all featured quite high human ground truth agreements, with 0.91, 0.89, and 0.98 respectively. In contrast, the clarity subset only had a score of 0.52, indicating almost random agreement. See tables 29, 30, 31, and 32 in Appendix A for a sample of each of the four subsets making up the RESPONSible Service Dataset.

3.2 Linguistic Data Profiling

The linguistic data profiling conducted on the RESPONSible Service dataset resulted in the compilation of a linguistic profile for each of the four featured styles, and was organized into six main sections: frequency, lexicon, discourse, deixis and modality, sentiment and emotion, and additional metrics. The analysis was carried out using a custom pipeline. All scores reported refer to averages across rows, with the exception of the proximal deictic ratio, which was computed at corpus level due to the low frequency of the relevant tokens. Statistical significance was computed for all analyses. The selection of test statistics was based on task type and data distribution. Chi-square tests were used for the frequency analysis. In the sentiment and emotion analysis, Rank-Sum tests were used to compare requests and responses within style-emotion pairs, while Kruskal-Wallis tests were applied to assess between-style statistical significance. Kruskal-Wallis tests were also used in all remaining analyses. Where Kruskal-Wallis tests were used, post hoc comparisons were conducted using Dunn’s test with Bonferroni correction. All test statistics used a significance threshold of $p < 0.05$. All analyses discussed in Sections 3.2 and 4.1 relate to comparisons among the four style subsets. The results can therefore be interpreted only in relative, not absolute, terms. Furthermore, the linguistic data profiling focused primarily on the `response1` column, and in some cases also included the `request` column. This choice was made to avoid unnecessary complexity by focusing on the more stylistically marked response type.

Frequency Analysis The first step of the analysis involved inspecting token and POS frequencies. As outlined in the literature review, the four styles, despite their overlap, are associated with distinct lexical tendencies. A frequency analysis can be helpful in identifying which tokens and POS tags occur more frequently within each style. For instance, one might expect the word “understand” to be more prevalent in the empathy subset, as it shows acknowledgment of the customer’s perspective (Clark et al., 2013). Frequency phenomena were investigated not only at the lexical level (individual words), but also at the morphosyntactic level (POS tags). For example, interjections, which contribute to create an informal atmosphere, might be more frequent in texts characterized by a friendly style (Ruytenbeek & Decock, 2024). Punctuation frequency was also analyzed as part of this section.

Lexical Analysis The lexical analysis focused on five dimensions: lexical diversity, lexical complexity, lexical repetition, information density, and text length. Lexical diversity was calculated using Type-Token Ratio (TTR), one of simplest and most widely used measures of lexical diversity (McCarthy & Jarvis, 2010). TTR simply measures how many unique words (types) are used relative to the total number of words (tokens), and ranges from 0 (low lexical diversity) to 1 (high lexical diversity). Lexical complexity was inspired by Maisto (2025) and operationalized as the proportion of lemmas falling within the first 2,000 items of the New General Service List (NGSL) (Browne et al., 2013), with higher proportions indicating lower lexical complexity. The NGSL is a frequency list that contains the 2,801 most frequent words in the English language as of the year 2013. Lexical repetition was also examined based on Maisto's (2025) approach by calculating repeated unigrams, bigrams, and trigrams within moving windows of 2, 5, and 10 tokens. Information density was defined as the ratio of content words to total words, based on Ure (1971). Finally, text length was measured simply as the number of tokens per response.

Discourse Analysis The discourse analysis concentrated on cohesion and relevance. Cohesion was assessed by quantifying the use of cohesive devices across additive, causal, temporal, and logical categories, based on a predefined list of 98 conjunctions and adverbs (see Appendix B), and normalized by text length, following Kormos's (2011) approach. Cohesion can improve clarity by signaling the relationships among parts of a text, thus creating a coherent and easily interpretable message. As we saw in Section 2, relevance, defined as the degree to which a response addresses the customer's needs, is also an important feature of CS communication, as it enhances clarity, empathy, and politeness (Lutzky, 2024; Packard & Berger, 2021). Irrelevant responses may indicate a lack of understanding or willingness to understand. Relevance was evaluated using ROUGE-L and BERTScore. ROUGE-L measures n-gram overlap based on the longest common subsequence (Lin, 2004), while BERTScore captures semantic similarity using word embeddings (Zhang et al., 2019). BERTScore is useful in those cases where two sentences overlap in meaning despite sharing little vocabulary. The use of both metrics ensured a balanced assessment of both surface-level and semantic overlap.

Analysis of Deixis and Modality Deictic and modal features were analyzed with a focus on linguistic elements known to influence customer perceptions. These elements are: proximal deictic ratio, pronoun type ratio, present tense ratio, and conditionals per text. The proximal deictic ratio was measured based on Zabava Ford (1999), who linked proximal deixis to verbal immediacy and friendliness. A curated list of deictic expressions (*here*, *there*, *this*, *that*, *these*, *those*) was used. Due to the low frequency of these tokens, this ratio was calculated at corpus level. Pronoun usage was analyzed by computing the ratio of first- and second-person pronouns, which have been shown to signal friendliness (Zabava Ford, 1999) and politeness (Friginal, 2024). Present tense ratio was also included, since the use of this tense, as opposed to past or future tenses, is associated with greater immediacy and engagement (Zabava Ford, 1999). Finally, conditional verbs per text were counted, as polite discourse often employs these linguistic devices to soften directness (Creelman, 2022).

Sentiment and Emotion Analysis Sentiment and emotion were also examined to account for the affective component of the discussed language styles, friendly and empathetic in particular. Sentiment and emotion were analyzed using the *twitter-roberta-base-sentiment* (Barbieri et al., 2020) and *emotion-english-distilroberta-base* (Hartmann, 2022) models, respectively. Emotion analysis was conducted to complement sentiment analysis, as sentiment can manifest through a wide range of emotions. Sentiment was classified as positive, neutral, or negative, and averaged across rows. The obtained scores were then normalized to be included between 0 and 1, with 1 being the most positive score. Similarly, the probability distributions across seven emotions (anger, disgust, fear, joy, neutral, sadness, and surprise) were computed and averaged. Sentiment and emotion analyses were applied to both customer requests

Frequency	Lexicon	Discourse	Deixis & Modality	Extra
Token frequency	Lexical diversity	Cohesive devices	Proximal deictic ratio	Sentiment
POS frequency	Lexical complexity	Response relevance	Pronoun type ratio	Emotion
Punctuation freq.	Lexical repetition		Present tense ratio	Readability
	Information density		Conditionals per text	Formality
	Text length			Concreteness

Table 1: Linguistic Data Profiling of the RESPONSible Dataset

and responses, reflecting the bidirectional nature of the analyzed styles, empathy in particular, in CS interactions (Ngo et al., 2020; Wieseke et al., 2012).

Additional Metrics Additional analyses were conducted to capture textual features not addressed by the aforementioned metrics. These included readability, formality, and concreteness. Readability was measured using the Flesch-Kincaid Reading Ease formula (Flesch, 1979), which measures textual complexity based on the number of syllables per word and words per sentence, and could thus be linked to clarity. Formality, a key dimension related to friendliness (Creelman, 2022; Ruytenbeek & Decock, 2024), was assessed using the *roberta-base-formality-ranker* (Babakov et al., 2023), with binary outputs averaged to produce a continuous score between 0 and 1, 0 being completely informal and 1 being completely formal. Finally, concreteness was inspired by Packard and Berger's (2021) findings and measured by adapting Maisto's (2025) method. Word concreteness scores were sourced from the MRC Psycholinguistic Database (Coltheart, 1981), which assigns each term a value between 100 and 700, 100 indicating the lowest level of concreteness and 700 the highest. The total concreteness scores of all identified words were summed and then normalized by dividing by the total number of tokens in each text multiplied by 100. Scores were also normalized to a [0,1] range, with higher values indicating greater concreteness. For a comprehensive overview of all analyses performed, please refer to Table 1.

3.3 Fine-tuning LLMs on Synthetic Data

The RESPONSible Service dataset was also evaluated in a text generation task carried out using three methods: in-context learning (ICL), supervised fine-tuning (SFT), and direct preference optimization (DPO). In this experiment, a selected LLM was given a CS message as input and prompted to generate an improved version of the message according to a specific language style (clear, friendly, empathetic, or polite), similarly to what was done with `response0` and `response1` when the RESPONSible Service dataset was first created. The generated output was then compared to the original dataset `response1` items, which acted as ground truth (GT). This experimental setting was chosen to enable the inclusion of the DPO setting, which requires two ranked items, one more desirable than the other. For this reason, unlike the linguistic data profiling, this experiment used the `response0` and `response1` columns. The model selected for the experiment was `unsloth/Llama-3.2-1B-Instruct` (Unsloth, 2024). This model was chosen for its small size, which makes it well-suited to SFT and DPO under limited computational resources. Since this experiment involved fine-tuning a smaller model (`Llama-3.2-1B-Instruct`) using data synthetically generated by a much larger model (`GPT-4o`), it can be characterized as an instance of *knowledge distillation* (Braga et al., 2024). The single experimental settings are described in the following paragraphs. All generations used `top-p` sampling (p = 1.0) and `temperature` set to 1.0. These parameters were specifically chosen to match those used during the original dataset generation, where they promoted greater diversity. Maintaining these settings ensured that the diversity of the outputs generated throughout the current task would be consistent with the diversity of the ground truth. Additionally, all generations were configured to produce a maximum of 40 new tokens, a value determined empirically to yield an average output length of approximately 30 tokens, thus matching the original dataset. Diagrams of the three experimental settings (ICL, SFT, and DPO) are provided in Figures 1, 2, and 3 respectively. The prompts and input formats used in all experimental settings have been included in Appendix C.

In-Context Learning (ICL) In the ICL setting, responses were generated using a Hugging Face pipeline. For each stylistic target, a separate model instance was prompted with the final 200 `response0` items from the ground truth (`GT-res0`, test). The model was conditioned on zero to five in-context examples, drawn randomly (with a fixed seed of 42) from the first 800 `response0`-`response1` pairs (`GT-res0`/`GT-res1`, train). Each prompt included a stylistically marked CS message and its stylistically enhanced version, formatted as: system instruction → base response → rewritten response. Prompts were constructed in advance and processed in batches (batch size = 8) to optimize GPU throughput. The stylistically modified counterparts in the outputs formed the `ICL-res1` (test) set.

Supervised Fine-Tuning (SFT) In the SFT setting, the same model architecture was fine-tuned using 800 training examples per style (`GT-res0`/`GT-res1`, train). Each training instance consisted of a structured chat message pairing a system instruction (requesting stylistic transformation) with a user input (`GT-res0`) and a supervised assistant reply (`GT-res1`). Tokenization applied a custom Unsloth-compatible chat template with a maximum sequence length of 2048. To reduce memory usage and accelerate training, parameter-efficient fine-tuning (PEFT) was employed using LoRA (Low-Rank Adaptation). Specifically, LoRA adapters with rank 64 and dropout of 0.1 were applied to the attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`). The model was fine-tuned with a learning rate of 2e-5 over a single epoch, using a batch size of 1 and gradient accumulation steps of 256. Gradient check-pointing and flash attention were enabled to further optimize memory efficiency. After training, the model was quantized to 4-bit NF4 precision with `bfloat16` compute type for inference. The model was then used to generate stylistically enhanced outputs for all items in `GT-res0` (train+test), resulting in the `SFT-res1` set.

Direct Preference Optimization (DPO) The DPO phase further refined the SFT stylistic outputs by applying pairwise preference training to the SFT-tuned models. For each style, the model previously fine-tuned via SFT was used as the initialization point. Training data consisted of response pairs: the SFT-generated outputs (`SFT-res1`, train) were treated as less preferred, while the corresponding ground truth targets (`GT-res1`, train) were labeled as preferred. These pairs were used to optimize the model using DPO, implemented via a single epoch of training with a batch size of 2, gradient accumulation steps of 4, and a learning rate of 5e-6. To improve efficiency, training was conducted using 4-bit quantization with the `adamw_8bit` optimizer. The model operated with either FP16 or BF16 precision depending on hardware support. A linear learning rate scheduler and a warmup ratio of 0.1 were applied, and DPO loss was controlled by a β value of 0.1. Training was performed without an explicit reference model, and each style's DPO-finetuned model was saved separately. For inference, the DPO-trained models were prompted with the test set from `GT-res0`, using a constrained chat template: each prompt included a system instruction requesting stylistic rephrasing followed by a user message (`GT-res0`). Tokenized prompts were batched and the generated outputs formed the final `DPO-res1` (test) set.

Baseline A baseline was also included using the non-fine-tuned model. This was loaded in 4-bit NF4 precision with `bfloat16` compute type for efficiency. Flash attention was enabled. The model was prompted with the same test inputs used in the SFT setting, using only the system messages (prompts) and user messages (`GT-res0`), thus excluding assistant replies. Tokenization applied a structured chat template. The resulting predictions formed the baseline set and served as a reference to assess the effects of stylistic supervision.

Model Evaluation The predictions generated in each experimental condition were compared to the ground truth (`GT-res1`) using the same two measures of overlap also featured in the linguistic data profiling: BERTScore for semantic overlap and ROUGE-L for formal (n-gram) overlap. See Section 3.2 for further details about these metrics. These measurements were carried out for all four styles. Two tables were produced to report BERTScores (F1) and ROUGE-L scores respectively. Statistical significance was assessed using the Kruskal-Wallis test, and Dunn's post hoc tests with Bonferroni correction were applied to identify pairwise differences between text generation settings. A difference was considered statistically significant when $p < 0.05$. A portion of the model predictions was also manually annotated by the author to assess human agreement. Specifically, 25 predictions were randomly sampled from each of the nine experimental conditions, resulting in 900 items in total (225 per style, 100 per condition). The process was the same used for the evaluation of the RESPONSible Service dataset (see Section 3.1). Results are reported in Section 4.2.

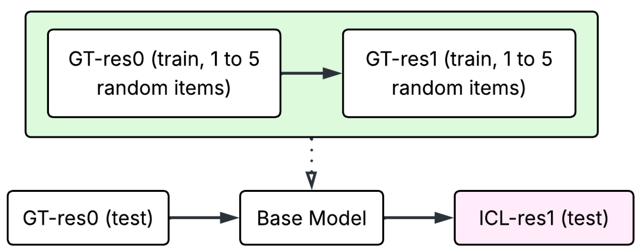

Figure 1: ICL Fine-tuning Scheme

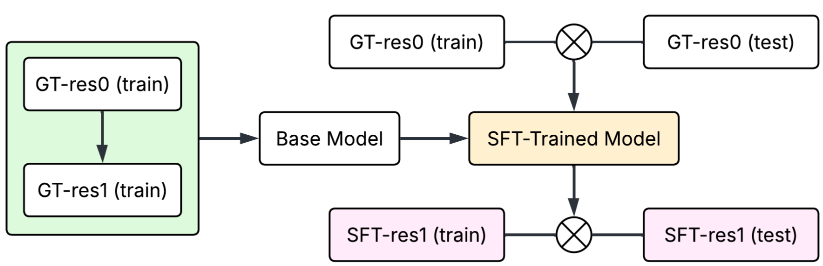

Figure 2: SFT Fine-tuning Scheme

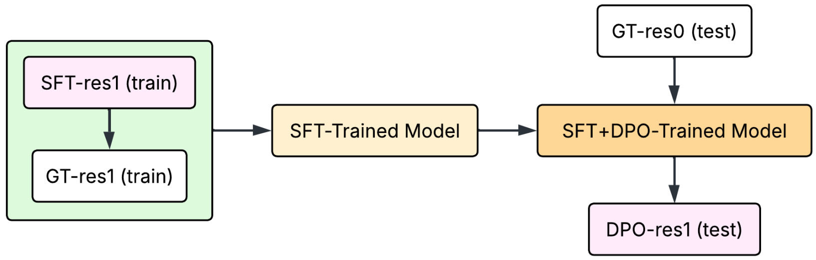

Figure 3: DPO Fine-tuning Scheme

4 Results and Discussion

4.1 Linguistic Data Profiling

In the presentation of all analyses, an indicator of statistical significance has been included in the `SS` (Statistical Significance) column. The possible `SS` labels are the following: an asterisk (`*`) indicates the presence of statistically significant differences with all other subsets; a hyphen (`-`) denotes the absence of statistically significant differences with any other subset; and, in cases of partial significance differences, the initials of the differing subsets are provided. For example, if an item in the clarity subset differs significantly only from the same item in the friendliness and politeness subsets, but not in the empathy subset, the label `fp` is used. For the specific test statistics used, see Section 3.2.

4.1.1 Frequency Analysis

Table 2 features the 20 most frequent tokens for each style subset, ranked by relative frequency. At the top of the table, it is evident that the word “thank” is the most frequent token across all subsets (4.5% for clarity, 3.5% for friendliness, and 3.7% for politeness) with the exception of empathy, where it appears only in 6th position, with a frequency of 1.8%. In the empathy subset, the top-ranked word is “truly” (2.7%), followed by “sorry” (2.4%) and “understand” (2.4%). The prominence of “truly” may reflect an effort by the CS agent to convey genuine and authentic engagement. It might be used to emphasize that apologies are sincerely felt and that there is a true intention to understand the customer’s situation. Further down the empathy list, “apologize” appears in 9th position with a frequency of 1.6%. Notably, “apologize” ranks higher in the politeness subset, occupying the 5th position at 2.3%. Although “sorry” and “apologize” convey similar meanings, “apologize” carries a more formal tone, which could be an indicator of negative politeness, as formality creates distance and thus reduces imposition. In the politeness subset, the 2nd most frequent word is “kindly” (3.5%). Commonly used to soften requests, “kindly” appears far less frequently in other subsets, surfacing only at the 20th position in the clarity subset (0.8%). Similarly, “sincerely” appears in both the politeness and empathy subsets, with a higher frequency in politeness (3.3%, 3rd position) compared to empathy (1.3%, 15th position). Although “sincerely” and “truly” are semantically close, their differing frequencies suggest that “sincerely” is more strongly associated with polite language. As Pinto (2011) notes, language that is traditionally associated with politeness often loses its original meaning through habitual use. It is plausible that the model interpreted “truly” as more heartfelt than “sincerely”, favoring its use in empathetic responses. Turning to the friendliness subset, the informal greetings “hi” and “hello” occupy the 2nd (2.6%) and 4th (2.0%) positions, respectively. No other subset features informal greetings among their 20 most frequent tokens, aligning with Zabava Ford’s (1999) concept of friendly service, where greetings play a key role in establishing rapport with customers. Additionally, “help” appears in 3rd position (2.5%), ranking higher than “assist”, which occupies the 9th position (1.3%). The greater frequency of “help”, a less formal alternative to “assist”, is consistent with the more informal tone of the friendliness subset. Conversely, in the politeness subset, while “assist” appears to be used with almost the same frequency as its counterpart in the friendliness subset (1.2% vs 1.3% respectively, no significant difference), “help” does not feature among the top twenty. In sum, the patterns identified in this section suggest that GPT-4o makes use of more or less evident lexical choices to distinguish the four language styles, as shown by the varying word distributions. At the same time, token frequency alone does not provide sufficient information to understand the way in which the model represents the four language styles, as many terms are used across them in similar ways.

#	**Clear**				**Friendly**				**Empathetic**				**Polite**			
	Token	Count	%	SS	Token	Count	%	SS	Token	Count	%	SS	Token	Count	%	SS
1	thank	698	4.5	*	thank	514	3.5	ce	truly	410	2.7	*	thank	609	3.7	ce
2	contact	394	2.5	*	hi	381	2.6	*	sorry	364	2.4	*	kindly	580	3.5	*
3	order	249	1.6	-	help	377	2.5	*	understand	363	2.4	*	sincerely	542	3.3	*
4	reach	232	1.5	ep	hello	302	2.0	*	resolve	332	2.2	*	appreciate	420	2.5	*
5	appreciate	222	1.4	ep	share	289	1.9	*	ensure	301	2.0	*	apologize	381	2.3	*
6	issue	211	1.3	fp	sorry	284	1.9	*	thank	273	1.8	*	inconvenience	342	2.1	*
7	detail	195	1.2	-	let	267	1.8	cp	let	261	1.7	cp	order	296	1.8	e
8	account	188	1.2	ep	order	263	1.8	e	inconvenience	247	1.6	*	share	268	1.6	*
9	number	183	1.2	*	assist	200	1.3	ce	apologize	238	1.6	*	contact	262	1.6	*
10	help	169	1.1	fp	reach	195	1.3	e	cause	230	1.5	*	promptly	244	1.5	*
11	refund	160	1.0	fp	contact	193	1.3	*	experience	220	1.5	*	patience	214	1.3	*
12	website	160	1.0	*	appreciate	187	1.3	p	order	203	1.4	fp	provide	209	1.3	*
13	provide	160	1.0	*	detail	186	1.3	-	share	201	1.3	*	detail	203	1.2	-
14	feedback	160	1.0	*	account	159	1.1	e	issue	198	1.3	fp	assist	198	1.2	ce
15	apologize	144	0.9	*	check	153	1.0	*	sincerely	195	1.3	*	customer	192	1.2	*
16	update	138	0.9	f	patience	148	1.0	cp	promptly	180	1.2	*	reach	183	1.1	ce
17	email	135	0.9	*	ensure	147	1.0	ce	important	166	1.1	*	cause	174	1.1	*
18	promptly	132	0.8	*	right	134	0.9	*	appreciate	163	1.1	cp	issue	158	1.0	*
19	support	130	0.8	*	quickly	131	0.9	cp	swiftly	162	1.1	*	ensure	154	0.9	ce
20	kindly	129	0.8	ep	number	127	0.9	ce	detail	161	1.1	-	inquiry	152	0.9	*

Table 2: Top 20 most frequent tokens by style subset

Table 3 presents all POS tags ranked by relative frequency for each style subset. Across all four subsets, the three most common parts of speech are nouns, verbs, and pronouns, although their order of prevalence varies by style. In the clarity subset, nouns are the most frequent (20.8%), followed by verbs (16.2%). This pattern may reflect the role of nouns and verbs as content words, which carry substantial informational weight. This aligns with one of the main criticisms in Lutzky’s (2024) study, where customers found Ryanair’s CS responses unclear due to their low information value. Although a higher proportion of nouns and verbs does not automatically ensure greater informativeness, their significant presence definitely goes in that direction. In contrast, the friendliness and empathy subsets prioritize pronouns, which are the most frequent POS categories (17.1% and 18.6%, respectively). Both styles aim to establish a personal connection with the interlocutor (Clark et al., 2013; Creelman, 2022; Ruytenbeek & Decock, 2024; Zabava Ford, 1999). This function is effectively fulfilled by personal pronouns. As we saw previously, first-person singular pronouns humanize the CS agent (Ruytenbeek & Decock, 2024), second-person pronouns add personalization by centering the customer’s perspective (Ruytenbeek & Decock, 2024), and first-person plural pronouns create a sense of inclusion (Zabava Ford, 1999). The politeness subset, like the clarity subset, shows a predominance of nouns (18.1%). However, pronouns (16.9%) are the 2nd most frequent POS, narrowly ahead of verbs (16.5%). Across all four subsets, punctuation consistently ranks 4th in frequency. Its use is highest in the friendliness subset (14.1%), followed by clarity (13.9%), politeness (11.8%), and empathy (10.6%). The prominence of punctuation in the friendliness subset aligns with Creelman’s (2022) findings, where punctuation is used to mimic the dynamics of spoken conversation. Punctuation may also be used to reinforce positive politeness, particularly through the use of exclamation marks (e.g., "Thank you for reaching out!") (Creelman, 2022). However, a closer examination of punctuation distribution (Table 4) reveals that exclamation marks account for 33.3% of punctuation in the friendliness subset, but only 1.6% in the politeness subset. This discrepancy suggests that the model associates politeness with negative politeness strategies more often than positive ones. The former indeed aim to minimize imposition and may therefore have led the model to use fewer exclamation marks. Additionally, the analysis shows that only in the friendliness subset is the exclamation mark (33.3%) used more frequently than the comma (14.9%), thus stressing the importance of exclamation marks in conveying friendliness, as well as suggesting a simpler and more direct writing style. Across the other subsets, exclamation mark usage is much lower, ranging from 4.1% in the clarity subset to 1.6% in politeness. Another interesting finding relates to the fact that the question mark appears least frequently in the clarity subset, possibly suggesting that the model interprets questions as introducing a degree of ambiguity that undermines clarity. However, the low statistical significance calls for caution in the interpretation of this result. Turning to other POS categories, interjections are more common in the friendliness subset (2.9%) than in clarity (1.5%), politeness (0.8%), and empathy (0.7%), again consistent with Creelman's (2022) observations. As for determiners, although these feature the highest relative frequency in both the clarity and empathy subsets, they rank higher in the former (6th position) than in the latter (9th position). This may reflect their contribution to precision and specificity, which are central to clear communication. Numerals are also more frequent in the clarity subset (0.9%) than in friendliness and politeness (0.4% each) or empathy (0.2%), suggesting that quantification could serve as another strategy for enhancing precision. Overall, analyzing POS frequency alongside token frequency gives us a clearer understanding of how GPT-4o selected words according to the required style. The way different parts of speech are used does not appear random, but partly reflects how the model interprets each language style dimension. The analysis also shows that punctuation, which is often overlooked in NLP, plays a meaningful role in shaping writing style because of its expressive power.

#	**Clear**				**Friendly**				**Empathetic**				**Polite**			
	POS	Count	%	SS	POS	Count	%	SS	POS	Count	%	SS	POS	Count	%	SS
1	NOUN	7471	20.8	*	PRON	6534	17.1	ce	PRON	6776	18.6	*	NOUN	6613	18.1	*
2	VERB	5825	16.2	f	VERB	5819	15.2	*	VERB	5837	16.0	f	PRON	6180	16.9	ce
3	PRON	5079	14.1	*	NOUN	5466	14.3	cp	NOUN	5409	14.8	cp	VERB	6043	16.5	f
4	PUNCT	4999	13.9	ep	PUNCT	5385	14.1	ep	PUNCT	3856	10.6	*	PUNCT	4321	11.8	*
5	ADP	3092	8.6	*	ADV	2942	7.7	c	AUX	2744	7.5	*	ADP	2881	7.9	cf
6	DET	1702	4.7	fp	ADP	2701	7.1	cp	ADP	2694	7.4	c	ADV	2726	7.4	c
7	ADJ	1568	4.4	ef	AUX	2288	6.0	*	ADV	2645	7.2	c	ADJ	1697	4.6	ef
8	AUX	1370	3.8	*	ADJ	1945	5.1	cp	ADJ	2005	5.5	cp	AUX	1697	4.6	*
9	ADV	1245	3.5	*	DET	1278	3.3	*	DET	1700	4.7	f	DET	1568	4.3	cf
10	CCONJ	1049	2.9	-	INTJ	1097	2.9	*	CCONJ	1097	3.0	f	CCONJ	1073	2.9	-
11	PART	744	2.1	ef	PART	1060	2.8	*	PART	890	2.4	*	PART	762	2.1	ef
12	INTJ	526	1.5	*	CCONJ	1021	2.7	e	SCONJ	459	1.3	fp	SCONJ	303	0.8	ce
13	PROPN	451	1.3	*	SCONJ	326	0.9	ce	INTJ	242	0.7	cf	INTJ	282	0.8	cf
14	SCONJ	417	1.2	fp	PROPN	228	0.6	ce	PROPN	85	0.2	*	PROPN	266	0.7	ce
15	NUM	341	0.9	*	NUM	161	0.4	ce	NUM	56	0.2	*	NUM	156	0.4	ce
16	SYM	42	0.1	*	SYM	19	0.0	ce	X	5	0.0	c	SYM	15	0.0	c
17	X	22	0.1	e	X	17	0.0	-	SYM	4	0.0	cf	X	13	0.0	-

Table 3: Frequency of POS tags by style subset

#	Clear				Friendly				Empathetic				Polite			
	Token	Count	%	SS	Token	Count	%	SS	Token	Count	%	SS	Token	Count	%	SS
1	.	3179	63.6	*	.	2331	43.3	*	.	2882	74.7	*	.	3078	71.2	*
2	,	1131	22.6	ef	!	1792	33.3	*	,	678	17.6	*	,	943	21.8	ef
3	!	205	4.1	*	,	804	14.9	*	-	114	3.0	c	-	96	2.2	cf
4	-	182	3.6	cp	-	161	3.0	cp	!	110	2.9	*	!	68	1.6	*
5	’	172	3.4	*	?	151	2.8	*	?	41	1.1	f	?	63	1.5	cf
6	”	62	1.2	*	’	65	1.2	ce	;	16	0.4	p	’	58	1.3	ce
7	%	43	0.8	*	”	34	0.6	*	’	8	0.2	*	%	18	0.4	c
8	?	37	0.7	fp	%	27	0.5	ce	%	6	0.2	cf	”	10	0.2	*

Table 4: Frequency of punctuation by style subset

4.1.2 Lexical Analysis

Table 5 presents the TTR scores for lexical diversity. All subsets exhibited high lexical diversity, with TTR scores falling between 0.865 and 0.871 and no significant differences between them. As for lexical complexity (Table 6), the clarity, friendliness, and empathy subsets exhibited higher proportions of high-frequency tokens (ranging from 0.68 to 0.69) than the politeness subset (0.59), suggesting that the model might represent polite language as relying more on less frequent, potentially more complex lexical items. In terms of lexical repetition (Table 7), the inter-style differences, which ranged between 5.92 for clarity and 5.56 for friendliness, were not found to be statistically significant. Moving to information density (Table 8), we can see that the politeness subset achieves the highest score (0.53), followed closely by the clarity subset (0.52), and by the friendliness and empathy subsets (both 0.49). This result is partly expected and partly counterintuitive. The clarity subset’s relatively high information density supports the notion that clear language should prioritize information delivery and minimizes formulaic expressions (Lutzky, 2024). This contrasts with friendly (Creelman, 2022; Ruytenbeek & Decock, 2024; Zabava Ford, 1999) and polite (Creelman, 2022; Hu et al., 2019; Pinto, 2011) language, which tend to incorporate more formulaic elements. The finding that friendly language has a lower density than clear language (0.49 vs. 0.52) aligns with this expectation. What is surprising is the score for the politeness subset, which is the highest. It suggests that the model either managed to maintain informativeness while embedding formulaic elements, or that it relied less on formulaicity in polite responses than in clear ones. Although this last hypothesis challenges prior findings regarding politeness in US culture, it is consistent with research indicating that formulaicity is not universally characteristic of politeness (Cameron, 1997; Friginal, 2024), and may, in some cultural contexts, even be perceived as impolite (Fuoli et al., 2021; Lutzky, 2024). Finally, the analysis of text length (Table 9) reveals one more pattern. The clarity subset has the shortest average text length (35.94 tokens), followed by the empathy (36.50) and politeness (36.60) subsets. The friendliness subset records the longest average length (38.29 tokens). This suggests a preference for conciseness in clear language, while friendly language is allowed to be more expressive, thus resulting in longer responses. This observation resonates with Zabava Ford’s (1999) findings, which highlight that friendly service communication not only conveys information but also promotes a welcoming atmosphere through phatic speech and verbal immediacy. In summary, these results show that some of the lexical patterns that are commonly found in human CS interactions according to previous literature are also present in LLM-generated ones. For instance, they show that GPT-4o associates clarity with conciseness, while politeness tends to be associated with a more complex vocabulary. However, part of the results indicate a divergence of the model from human lexical patterns. In particular, the unexpectedly high information density in the politeness subset suggests that GPT-4o does not necessarily express politeness through a reduction of general informativeness.

Subset	Score	SS
clear	0.870	-
friendly	0.871	-
empathetic	0.865	-
polite	0.865	-

Table 5: Lexical Diversity (TTR, average)

Subset	Score	SS
clear	0.69	p
friendly	0.69	p
empathetic	0.68	p
polite	0.59	*

Table 6: Lexical Complexity (NGSL frequency list, average)

Subset	Score	SS
clear	5.92	-
friendly	5.56	-
empathetic	5.76	-
polite	5.73	-

Table 7: Lexical Repetition (across n-grams, average)

Subset	Score	SS
clear	0.52	*
friendly	0.49	cp
empathetic	0.49	cp
polite	0.53	*

Table 8: Information density (content words / total words, average)

Subset	Score	SS
clear	35.94	*
friendly	38.29	*
empathetic	36.50	cf
polite	36.60	cf

Table 9: Text length (average)

4.1.3 Discourse Analysis

Regarding the use of cohesive devices (Table 10), the clarity subset features the highest ratio, with 5% of tokens per text being cohesive devices on average. All other subsets report a lower average of 4%. Importantly, only the difference between the clarity subset and the remaining subsets was found to be statistically significant. This suggests a meaningful distinction in the use of cohesive devices between clear language and the other analyzed styles. The model appears to associate clarity with a more frequent use of cohesive elements, which aligns with their recognized role in human-authored texts. Indeed, cohesive devices function as structural connectors that facilitate smoother transitions between textual segments, thus enhancing comprehensibility.

The analysis then considered the semantic and n-gram overlap between customer requests and CS responses. Beginning with semantic similarity, assessed via BERTScore (Table 11), the empathy subset stands out as the only condition that differs significantly from the others. It also records the lowest average score (0.54), compared to the 0.55 score shared by the other subsets. This indicates that request–response pairs in the empathy subset exhibit less semantic overlap, suggesting a weaker alignment in meaning. This result appears to diverge from established findings in human CS communication. Specifically, prior research has shown that customers perceive responses as more empathetic when they feel heard as a consequence of CS responses being closely aligned with their requests (Clark et al., 2013; Packard & Berger, 2021; Tan et al., 2019). One interpretation is that the model's representation of empathy may not be rooted in semantic alignment or direct relevance, but rather in other linguistic or stylistic cues. In contrast, the analysis of n-gram overlap using ROUGE-L (Table 12) reveals a different trend. Here, the empathy subset achieves the highest score (0.18), followed by the clarity and friendliness subsets (both 0.17), which do not differ significantly. The politeness subset ranks lowest with a score of 0.16. Unlike the semantic overlap results, these findings align more closely with previous observations (Clark et al., 2013; Packard & Berger, 2021; Tan et al., 2019). The apparent discrepancy between BERTScore and ROUGE-L highlights a distinction between semantic and surface-level similarity. While empathetic responses appear to retain much of the lexical content of the original request, thus explaining their higher ROUGE-L score, they may simultaneously introduce subtle semantic shifts that BERTScore, due to its reliance on contextual embeddings, is more sensitive to. This suggests that the model's empathetic responses prioritize lexical mirroring over deeper semantic congruence. It is possible that this behavior is common even in human CS communication, which in turn might have affected the model's underlying representations of empathetic communication.

Subset	Ratio	SS
clear	0.05	*
friendly	0.04	c
empathetic	0.04	c
polite	0.04	c

Table 10: Cohesion (ratio of connectives, average)

Subset	Score	SS
clear	0.55	e
friendly	0.55	e
empathetic	0.54	*
polite	0.55	e

Table 11: Semantic req-res overlap (BERTScore (F1), average)

Subset	Score	SS
clear	0.17	ep
friendly	0.17	ep
empathetic	0.18	*
polite	0.16	*

Table 12: N-gram req-res overlap (ROUGE-L, average)

4.1.4 Analysis of Deixis and Modality

The analysis of proximal deictics revealed statistically significant differences across all subsets (see Table 13). The empathy subset exhibited the highest proportion of proximal deictics relative to total deictics, with 66% of deictics used being proximal. This was followed by the friendliness subset (37%), the politeness subset (29%), and finally the clarity subset, which recorded the lowest proportion at 20%. The model's preference for proximal deictics in the empathy subset can be interpreted in light of Packard and Berger's (2021) findings. Proximal deictics, which refer to entities nearer in space or time (e.g., "this", "here", "now"), may be perceived as more concrete and specific than their distal counterparts. Their use may signal attentiveness to the immediate context of the customer, thus improving the perceived empathy of the response. In particular, the empathy subset is the only condition in which proximal deictics constitute the majority, suggesting a deliberate stylistic strategy by the model to highlight contextual closeness in empathetic communication. Conversely, the clarity subset shows the lowest use of proximal deictics. This result suggests that clarity, as modeled here, may prioritize referential clarity through alternative strategies rather than through deictic proximity. The friendliness subset, while not matching the empathy subset in proximal usage, still ranks 2nd. This finding is consistent with Zabava Ford's (1999) concept of verbal immediacy, which plays a key role in promoting a sense of intimacy and approachability in service interactions. Such immediacy can contribute to the friendly tone by reducing perceived distance between the customer and the agent.

The model's use of the present tense, as shown in Table 14, generally reflects patterns found in earlier research. The friendliness subset shows the highest average ratio of present tense verbs to total verbs, at 94%. This is again consistent with Zabava Ford's (1999) verbal immediacy: like proximal deictics, the present tense, compared to the past and the future, refers to moments that feel closer to the speaker and listener, making the interaction seem more immediate and personal. Following friendliness, the clarity and empathy subsets both record a ratio of 91%, while the politeness subset scores lowest at 89%. The slightly lower values for clarity and empathy do not have an immediately clear explanation. However, the even lower score in the politeness subset may be linked to the use of negative politeness strategies, which aim to reduce imposition and are often expressed through more indirect language (Brown & Levinson, 1987; Creelman, 2022; Hu et al., 2019). This tendency is especially common in cultures like the British one, where negative politeness is emphasized (Cameron, 1997). From this perspective, the present tense may be seen as too direct, since it anchors the message to the immediate moment and can reduce the sense of distance between speaker and listener. It is also worth noting that all four subsets rely heavily on the present tense overall. This suggests that, compared to proximal deictics, the model may treat present tense as a more reliable marker of effective CS communication, perhaps because it signals attentiveness and relevance without being overly specific or context-dependent.

Turning to the use of conditionals per text, a clear divide emerges between the clarity and empathy subsets on one side, and the friendliness and politeness subsets on the other. The clarity and empathy subsets feature relatively low averages, with 0.09 and 0.10 conditionals per text respectively, while the friendliness and politeness subsets record higher averages of 0.23 and 0.24 (see Table 15 for the complete results). These findings are consistent with earlier observations. The lower use of conditionals in the clarity subset can be explained by the emphasis that clear language places on certainty and directness. In contrast, conditionals tend to introduce uncertainty. Similarly, drawing on Packard and Berger's (2021) findings, it can be argued that perception of empathy is tied to the use of concrete and specific language. Since conditionals often express hypothetical or uncertain ideas, they may reduce the sense of concreteness that supports said perception. On the other hand, the higher frequency of conditionals in polite language fits well with previous research on negative politeness strategies, which promote indirectness as a way to soften potential imposition (Brown & Levinson, 1987; Creelman, 2022; Hu et al., 2019). Creelman (2022) specifically identifies the use of conditionals as a marker of polite speech. The friendliness subset's higher use of conditionals is more surprising. It appears to conflict with Zabava Ford's (1999) notion of verbal immediacy, which is central to the idea of friendly and courteous service. It also contrasts with the association between friendliness and the use of colloquial and informal language (CHV), as discussed by Creelman (2022), and Ruytenbeek and Decock (2024), which typically aims to reduce social distance. One possible explanation is that the model may have conflated friendliness with related traits such as niceness or kindness, which often encourage a more cautious and softened way of speaking. This overlap between categories could account for the unexpected results observed in the friendliness subset.

As for personal reference, the distribution of first-person and second-person pronouns is reported in Tables 16, 17, and 18. Regarding first-person plural pronouns, the politeness and friendliness subsets show the highest ratios, at 45% and 44% respectively, with no statistically significant difference between them. These are followed by the clarity subset (42%), which differs significantly from politeness but not from friendliness. The empathy subset records the lowest ratio at 32%, and it is the only subset that differs significantly from all others. Turning to second-person pronouns, the empathy and friendliness subsets score lowest, both at 50%. These are followed by the politeness subset (53%) and the clarity subset (56%). The analysis of first-person singular pronouns shows a different pattern. Here, all subsets differ significantly from one another. The empathy subset records the highest ratio (17.4%), followed by friendliness (5.5%), politeness (2.1%), and clarity (0.7%). Overall, the analysis of pronoun use only partially aligns with previous findings. The politeness subset, for instance, features a higher proportion of second-person pronouns than the empathy and friendliness subsets. While these findings are consistent with observations by Friginal (2024), who attribute second-person pronouns to polite language, they are also surprising. Indeed, one might expect second-person pronouns to appear more frequently in empathetic language, where they help center the customer's perspective (Ruytenbeek & Decock, 2024), than in polite language, where this type of pronoun might be seen as too direct. This latter intuition is corroborated by the politeness subset featuring fewer second-person pronouns than the clarity subset. Indeed, clarity favors more direct language, and second-person pronouns provide a straightforward way of addressing the customer compared to more indirect constructions, such as impersonal phrases or passive voice, which may be preferred in polite language. Interestingly, the clarity subset shows the lowest use of first-person singular pronouns. This may suggest that such pronouns are less effective in establishing clear communication, as they center the message on the speaker's perspective rather than the listener's. The politeness subset's high ratio of first-person plural pronouns is also notable. However, it is important to distinguish between inclusive and exclusive uses of the first-person plural. If the pronouns are inclusive ("we" meaning both the agent and the customer), this would seem to contradict the indirectness expected of polite language, as it reduces the perceived distance between the two parties. If instead the pronouns are exclusive ("we" referring only to the agent's company, for instance), this would align better with the distancing function associated with politeness. An in-depth qualitative analysis would be needed to clarify which interpretation is correct. In addition, politeness shows a low ratio of first-person singular pronouns compared to other subsets. One possible explanation is that first-person singular references might present the speaker as overly self-focused, which would be in contrast with the goal of minimizing imposition in negative politeness. For the friendliness subset, the results on first-person pronouns generally align with Ruytenbeek and Decock's (2024) observation that such pronouns can highlight "the human beings behind the professional/commercial roles". However, the lower use of second-person pronouns is less expected. Although second-person pronouns are often used to personalize communication and create a conversational tone, their lower presence here may again reflect a conflation of friendliness with politeness, leading the model to avoid forms of address perceived as too direct. Finally, the empathy subset presents an interesting profile. It records the lowest ratios of both first-person plural and second-person pronouns, but the highest use of first-person singular pronouns. This may be related to the frequent use of apologies and expressions of understanding, such as "I am so sorry" or "I understand how frustrating this must be". However, as previously mentioned, it remains surprising that empathetic responses do not feature more references to the customer's experience through second-person pronouns, or attempts to create a shared perspective through inclusive first-person plural forms. Although prior literature does not explicitly address pronoun use in empathetic communication, one might expect the use of "you" and "we" to be more effective in conveying empathy compared to "I".

In sum, the analysis of deictics and modality integrates previous analyses providing us with additional insights into the way GPT-4o represents language style. The analysis reveals that many stop words that are commonly removed in text preprocessing, such as deictics, conditionals, and pronouns, as well as implied grammatical features such as verb tense, rather than being noise, can contribute meaningfully to the stylistic nuances of model-generated texts.

Subset	Ratio	SS
clear	0.20	*
friendly	0.37	*
empathetic	0.66	*
polite	0.29	*

Table 13: Proximal deictic ratio (corpus-level)

Subset	Ratio	SS
clear	0.91	fp
friendly	0.94	*
empathetic	0.91	fp
polite	0.89	*

Table 14: Present tense ratio (average)

Subset	Count	SS
clear	0.09	fp
friendly	0.23	ce
empathetic	0.10	fp
polite	0.24	ce

Table 15: Conditionals per text (average)

Subset	Ratio	SS
clear	0.007	*
friendly	0.055	*
empathetic	0.174	*
polite	0.021	*

Table 16: Pronoun type ratio (1S person, average)

Subset	Ratio	SS
clear	0.42	ep
friendly	0.44	e
empathetic	0.32	*
polite	0.45	ce

Table 17: Pronoun type ratio (1P person, average)

Subset	Ratio	SS
clear	0.56	*
friendly	0.50	cp
empathetic	0.50	cp
polite	0.53	*

Table 18: Pronoun type ratio (2 person, average)

4.1.5 Sentiment and Emotion Analysis

The sentiment analysis results are presented in Tables 19 and 20. We can see how the ranking order is the same for both requests and responses: friendliness scores highest (0.40 for requests, 0.87 for responses), followed by politeness (0.38 and 0.81), clarity (0.37 and 0.75), and empathy (0.34 and 0.65). It is important to note that, overall, customer requests show much lower sentiment scores compared to the responses. This difference, which was statistically tested using the Rank-sum test (see Section 3.2), was found to be significant. In all cases, request sentiment remains below 0.50, while response sentiment consistently exceeds this threshold. Additionally, there is less variation among request scores (a range of 0.06), with only the empathy subset differing significantly from friendliness and politeness. These results are expected, as customers' requests were not generated according to specific style guidelines. By contrast, the responses show a broader range (0.22) and statistically significant differences across all subsets, reflecting the model's stylistic adjustments. At any rate, these findings offer a quite clear overview of how the model represents sentiment in CS interactions. Model-generated customer requests generally feature lower sentiment, likely reflecting the frustration or concern that commonly prompts a service request. In response, the model generates replies that are significantly more positive, as if to lift the general mood of the exchange. In addition to this, the style of the response influences the degree of sentiment adjustment: friendliness leads to the largest increase in positivity (47 points), likely due to the use of positive intensifiers, interjections, exclamation marks, and emotionally charged words such as "great" and "excited" (Creelman, 2022). At the other end of the spectrum, empathetic responses show the smallest sentiment increase (31 points). This is consistent with the idea that empathetic communication involves aligning more closely with the customer's emotional state, thus resulting in a less positive shift. Finally, the fact that both requests and responses share the same ranking in terms of sentiment aligns with prior findings regarding the reciprocal nature of many communicative styles (Hu et al., 2019; Ngo et al., 2020; Wieseke et al., 2012). Indeed, lower sentiment in the requests was met with lower sentiment in the responses, and vice versa.

A similar analysis was conducted for emotion, with results reported in Tables 21 and 22. In this analysis, the harmonic mean `HM` of emotions across styles was also included. Statistical differences between customer requests and CS responses were also assessed using the Rank-sum test. A statistically significant difference was found between requests and responses, except for joy in empathy and sadness in politeness. As with sentiment, style appears to influence emotional differences across the subsets. In particular, Table 22 (responses) shows a larger number of statistically significant differences than Table 21 (requests). This aligns with the observations made in the sentiment analysis, where sentiment was more uniform in requests than in responses. In general, all emotions except for joy are more strongly represented in requests than in responses. Neutral and sadness are the most common emotions in both requests (neutral: 0.403; sadness: 0.291) and responses (neutral: 0.280; sadness: 0.286). In customer requests, surprise is the third most represented emotion (0.130), while in responses, it is joy (0.173). Focusing on the responses, joy (0.119) and surprise (0.068) are most prominent in the friendliness subset. The positive association between friendliness and joy is consistent with the findings of the sentiment analysis, where friendliness emerged as the most positive style. This result may again reflect the use of positive language identified by Creelman (2022) as a feature of friendly communication. The clarity subset, on the other hand, features the highest proportion of neutral responses (0.578). This suggests that the model may have deliberately minimized emotional expression in clear communication, aligning with the idea that strong emotions, which have a significant subjective component, could undermine the objectivity associated with clarity. Negative emotions, namely sadness (0.493), fear (0.057), and anger (0.048), are most strongly represented in the empathy subset. This finding also mirrors the sentiment analysis, in which empathy was associated with lower positivity. It seems likely that the model attempted to convey empathy by reflecting the emotional tone of the customers' requests, which were more often neutral or negative. Finally, an interesting pattern emerges in the politeness subset, where disgust, although rare overall, features the largest representation (0.012) compared to the other three styles. One possible explanation is that the emotion classifier may have associated the more distant and polished tone of negative politeness with snobbishness, which could be associated with a subtle form of disgust.

In summary, the sentiment and emotion analysis reveals that even less strictly linguistic features, such as sentiment polarity and emotion, play a central role in how GPT-4o differentiates between styles. The model consistently produces responses that are more positive than the corresponding customer inputs, but this positivity is adapted according to the model's stylistic goals: friendliness amplifies joy and sentiment, while empathy preserves alignment with the customer's negative emotional state. These findings highlight that sentiment and emotion are integral components of the model's underlying stylistic representations.

Subset	**Score**	**SS**
clear	0.37	-
friendly	0.40	e
empathetic	0.34	fp
polite	0.38	e

Table 19: Sentiment Score (requests, average)

Subset	**Score**	**SS**
clear	0.75	*
friendly	0.87	*
empathetic	0.65	*
polite	0.81	*

Table 20: Sentiment Score (responses, average)

Emotion	**Clear**	**SS**	**Friendly**	**SS**	**Empath.**	**SS**	**Polite**	**SS**	**HM**
anger	0.035	e	0.028	e	0.048	*	0.036	e	0.035
disgust	0.012	-	0.010	-	0.012	-	0.012	-	0.011
fear	0.051	-	0.050	-	0.065	-	0.052	-	0.054
joy	0.070	f	0.087	ce	0.060	fp	0.069	e	0.070
neutral	0.427	ef	0.392	cp	0.374	cp	0.423	ef	0.403
sadness	0.278	e	0.295	-	0.315	cp	0.278	e	0.291
surprise	0.127	-	0.138	e	0.125	f	0.130	-	0.130

Table 21: Emotion Score (requests, average)

Emotion	**Clear**	**SS**	**Friendly**	**SS**	**Empath.**	**SS**	**Polite**	**SS**	**HM**
anger	0.018	*	0.016	cp	0.048	cp	0.017	*	0.020
disgust	0.010	fp	0.006	*	0.009	fp	0.012	*	0.009
fear	0.031	ef	0.037	*	0.057	*	0.055	ef	0.042
joy	0.119	*	0.433	*	0.130	*	0.214	*	0.173
neutral	0.578	*	0.174	*	0.241	*	0.377	*	0.280
sadness	0.204	ep	0.266	ep	0.493	*	0.305	*	0.286
surprise	0.039	*	0.068	*	0.022	*	0.020	*	0.030

Table 22: Emotion Score (responses, average)

4.1.6 Additional Metrics

The results of the formality analysis, presented in Table 23, are straightforward but informative. Although the formality classifier used in this study was not particularly sensitive to subtle differences, a clear pattern emerges. The clarity, empathy, and politeness subsets all feature an average formality score of 1.00, meaning that all texts in these subsets were classified as formal. In contrast, the friendliness subset, the only subset to differ significantly from the others, records a slightly lower average formality score of 0.95. This finding supports the link between friendliness and informality discussed in previous literature (Creelman, 2022).

Turning to readability, the results are summarized in Table 24. The friendliness subset achieved the highest Flesch Reading Ease score (69.72), followed by clarity (56.32), empathy (52.80), and politeness (45.31). The high readability score for friendly language can again be linked to its association with informal communication (Creelman, 2022). In contrast, the politeness subset's lower score aligns with the findings of the earlier analysis of lexical complexity. As we saw previously (see Section 4.1.2), politeness was found to feature more complex words and sentences, which in turn can reduce readability. Although clarity did not achieve the highest readability score, it ranked second, which supports the idea that more readable texts, with shorter sentences and simpler words, enhance clarity.

Finally, the analysis of language concreteness, reported in Table 25, reveals a pattern similar to the one observed in readability. The friendliness subset scores highest for concreteness (0.287), followed by empathy (0.281), clarity (0.274), and politeness (0.265). It is plausible that the informal and simpler style of friendly language leads to more concrete references, as informal speech often deals with more easily understandable everyday topics. The result for the empathy subset is also noteworthy. The second position in concreteness aligns with Packard and Berger's (2021) finding that more concrete language improves the perception of empathy and contributes to greater customer satisfaction.

Subset	**Score**	**SS**
clear	1.00	f
friendly	0.95	*
empathetic	1.00	f
polite	1.00	f

Table 23: Formality Score (average)

Subset	**Score**	**SS**
clear	56.32	*
friendly	69.72	*
empathetic	52.80	*
polite	45.31	*

Table 24: Readability Score (Flesch Reading Ease, average)

	Score	SS
Subset	Score	SS
clear	0.274	*
friendly	0.287	*
empathetic	0.281	*
polite	0.265	*

Table 25: Concreteness Score (average)

4.2 Fine-tuning LLMs on Synthetic Data

Both BERTScore and ROUGE-L scores for the text generation task can be found in Tables 26 and 27. The `SS` column uses the same system introduced in the linguistic data profiling to indicate statistical significance (see Section 4.1). However, in this case, the `SS` symbols indicate differences between text generation settings and not between subsets. The alphanumeric symbols to indicate partially significant differences are the following: digits from `0` to `5` represent 0- to 5-shot ICL, `b` for the baseline, `s` for SFT, and `d` for DPO. For the specific test statistics used, see Section 3.3. Unless specified otherwise, metric results are presented with the BERTScore first and ROUGE-L following. The harmonic means (`HM`) of both metrics across styles were also included in the tables.

Starting from the clarity subset, the highest overlap was observed in the 1-shot ICL setting (0.68 and 0.35). All ICL variants performed similarly (0.66–0.68, 0.31–0.35), clearly outperforming the baseline and SFT (both 0.59, 0.26), as well as DPO (0.57, 0.20). In friendliness, ICL again showed the strongest overlap, with 0.62–0.63 and 0.25–0.27. These were higher than the scores of the baseline (0.55, 0.20), SFT (0.56, 0.22), and DPO (0.53, 0.18). The empathy subset followed the same trend. ICL achieved 0.61–0.62 and 0.23–0.25, ahead of the baseline (0.57, 0.21), SFT (0.56, 0.21), and DPO (0.55, 0.19). Finally, in politeness, ICL once again led with scores of 0.63–0.64 and 0.25–0.27. These were followed by the baseline (0.59, 0.21), SFT (0.60, 0.22), and DPO (0.56, 0.19).

Overall, as we can see, ICL exhibited the highest degree of overlap with the ground truth across all styles, with 0.63 and 0.26–0.28. Furthermore, overlap scores were generally consistent across ICL settings, except in the clarity subset, where greater variability was observed. This fluctuation may reflect the less stylistically predictable nature of the clarity subset, as indicated by the lower human agreement discussed in Section 3.1. As for the baseline and SFT, these generally featured lower ground truth overlap than ICL (0.58, 0.22–0.23), with no significant differences observed between them in any style. Finally, DPO consistently showed the lowest ground truth overlap (0.55, 0.19) and was frequently outperformed by most or all of the other conditions. It is important to note that a lower n-gram overlap with the GT does not necessarily indicate poor data quality, since text generation tasks often allow for multiple valid outputs for the same input. However, when this lower overlap is accompanied by a drop in semantic similarity, it may signal deeper issues. In tasks like text style transfer, only the style should be modified, while the core meaning should remain unchanged.

Exp. Setting	**Clarity**		**Friendliness**		**Empathy**		**Politeness**		**HM**
	F1	SS	F1	SS	F1	SS	F1	SS	
ICL (0-shot)	0.66	1bds	0.62	bds	0.61	bds	0.63	bds	0.63
ICL (1-shot)	0.68	025bds	0.62	bds	0.61	bds	0.64	bds	0.63
ICL (2-shot)	0.66	1bds	0.62	bds	0.62	bds	0.64	bds	0.63
ICL (3-shot)	0.66	bds	0.62	bds	0.61	bds	0.63	bds	0.63
ICL (4-shot)	0.66	bds	0.62	bds	0.61	bds	0.64	bds	0.63
ICL (5-shot)	0.66	1bds	0.63	bds	0.61	bds	0.64	bds	0.63
Baseline	0.59	012345d	0.55	012345	0.57	012345d	0.59	012345	0.58
SFT	0.59	012345d	0.56	012345d	0.56	012345	0.60	012345d	0.58
DPO	0.57	*	0.53	012345s	0.55	012345b	0.56	012345s	0.55

Table 26: BERTScore (F1) across Styles (Llama-3.2-1B-Instruct)

Exp. Setting	**Clarity**		**Friendliness**		**Empathy**		**Politeness**		**HM**
	R-L	SS	R-L	SS	R-L	SS	R-L	SS	
ICL (0-shot)	0.32	bds	0.25	bds	0.25	bds	0.27	bds	0.27
ICL (1-shot)	0.35	5bds	0.26	bds	0.23	d	0.27	bds	0.28
ICL (2-shot)	0.32	bds	0.25	bds	0.25	bds	0.27	bds	0.27
ICL (3-shot)	0.31	bds	0.26	bds	0.23	bds	0.25	bds	0.26
ICL (4-shot)	0.32	bds	0.27	bds	0.24	bds	0.26	bds	0.27
ICL (5-shot)	0.31	1bds	0.27	bds	0.24	bds	0.27	bds	0.27
Baseline	0.26	012345d	0.20	012345	0.21	02345	0.21	012345	0.22
SFT	0.26	012345d	0.22	012345d	0.21	02345	0.22	012345d	0.23
DPO	0.20	*	0.18	012345s	0.19	012345	0.19	012345s	0.19

Table 27: ROUGE-L Scores across Styles (Llama-3.2-1B-Instruct)

Turning to human agreement (see Table 28), when comparing the results across subsets, a clear distinction emerges: the agreement rates for clarity were consistently lower than those for the other styles, with an average of 30% compared to 50% for friendliness, 66% for empathy, and 59% for politeness. This pattern suggests that the inconsistencies in stylistic marking observed in the clarity subset of the RESPONSible Service dataset (see Section 3.1) carried over to the outputs of models trained on it, making the more stylistically marked responses harder to identify than in other styles. Human agreement also differed notably between the ICL settings and the other conditions (baseline, SFT, and DPO). ICL consistently received higher agreement scores, averaging between 54% and 61%, whereas the baseline and SFT averaged just 22% and 21%, respectively. Human agreement scores in the 0-shot ICL setting do not appear to have been negatively affected by the absence of examples from the RESPONSible Service dataset. In fact, they were even higher compared to the ones for the few-shot ICL. This aligns with previous findings, which show that ICL can sometimes perform better with irrelevant or no context than with relevant context (Li et al., 2024). DPO, in turn, performed better than the baseline and SFT, reaching up to 47% agreement, but it still fell short of ICL performance.

Exp. Setting	**Clarity**	**Friendliness**	**Empathy**	**Politeness**	**HM**
ICL (0-shot)	36%	68%	88%	76%	60%
ICL (1-shot)	32%	64%	76%	68%	54%
ICL (2-shot)	36%	60%	80%	76%	57%
ICL (3-shot)	44%	60%	80%	72%	61%
ICL (4-shot)	36%	60%	84%	72%	57%
ICL (5-shot)	40%	64%	76%	68%	59%
Baseline	16%	24%	32%	20%	22%
SFT	12%	32%	20%	36%	21%
DPO	36%	48%	56%	48%	47%
HM	30%	50%	66%	59%	–

Table 28: Human agreement (%) across Styles (Llama-3.2-1B-Instruct)

Overall, the analysis of n-gram overlap, semantic similarity, and human agreement consistently indicates that ICL had an advantage over both SFT and DPO. One possible explanation is the quality of the GT data. Specifically, the lower overlap between model outputs and GT after training on the RESPONSible Service dataset, compared to when few or no examples were provided, may suggest that the dataset contains noise or lacks representativeness. Wang et al. (2023) found that ICL is generally more robust to label noise than SFT in text classification tasks. Although text generation differs from text classification, the possibility that this difference in robustness also applies to generative settings could help explain the observed results.

5 Conclusion

The main goal of this study was to explore the intersection of AI-based text generation, stylistic variation, and CS communication. By generating a synthetic dataset (RESPONSible Service) using GPT-4o, I investigated whether and how LLMs can represent and distinguish key CS communication styles: clarity, friendliness, empathy, and politeness. This study explored four main research questions: what language patterns emerge in LLM-generated CS interactions (RQ1), how these patterns vary across different CS communication styles (RQ2), how they compare to those discussed in prior research on human CS communication (RQ3), and in what way the text generation performance of models adapted to LLM-generated CS data can be evaluated and interpreted (RQ4). The results provide insights into both the capabilities and limitations of LLMs in the field of CS, and into their linguistic representations of CS-related language style.

5.1 Main Findings

Analysis of the dataset showed that GPT-4o produced systematic linguistic variation across styles. Each style was distinguished by characteristic patterns across lexical, morphosyntactic, semantic, and discourse levels. Clarity was marked by shorter texts, greater use of cohesive devices, determiners, second-person pronouns, and emotionally neutral words, as well as limited use of conditionals. This partially align with Zabava Ford's (1999) suggestion that clear communication is based on the use of simple and straightforward language. Friendliness featured more informal, expressive, and engaging language. It scored lowest in formality, and highest in text length, present-tense usage, sentiment, and joy. Furthermore, it featured more phatic markers, especially greetings such as "hi" and "hello", as well as interjections and punctuation, exclamation marks in particular. All these features are in line with Zabava Ford's (1999) notion of verbal immediacy and with the concept of CHV (Creelman, 2022; Ruytenbeek & Decock, 2024), both of which help reduce social distance and make interactions more personable. The empathy subset revealed that GPT-4o tends to convey this style primarily through surface-level features, such as increased usage of words like "truly" and "understand", and higher n-gram overlap between customer requests and CS responses. However, this seemed to happen without an actual understanding of the customer's situation, as shown by the low semantic overlap. This could be seen as contrasting with Clark et al.'s (2013) idea of cognitive empathy, where true understanding and alignment with the customer's situation and emotions is key in effective CS. The LLM's behavior suggests that surface-level linguistic empathy features might be common in human CS communication, which in turn might have influenced the way LLMs interpret empathetic CS communication. Empathy as modeled by GPT-4o also featured the highest ratio of proximal deictics, first-person singular pronouns, and negative emotions such as sadness, fear, and anger, as well as the lowest ratio of first-person plural pronouns and sentiment score. In particular, empathy's sentiment and emotion profile aligned with previous findings about the reciprocal nature of this communicative style (Ngo et al., 2020; Wieseke et al., 2012), as in this case the negative sentiment and emotions of the customer seemed to have influenced CS responses. Finally, politeness involved a greater use of formal words, such as "apologize" in place of "sorry" or "assist" in place of "help", as well as of linguistic forms indicating indirectness, such as conditionals and non-present verb forms. Politeness also featured higher lexical complexity. Although polite language is often formulaic in nature, GPT-4o still represented it featuring high information density, suggesting that the LLM can balance formulaicity and informativeness in ways that might be less common in human communication.

To test the dataset's quality and usability in text generation tasks, models were trained or prompted to stylistically improve `GT-res0` items and evaluated against `GT-res1` using BERTScore and ROUGE-L, measures of semantic and formal overlap respectively, as well as against human judgments. Overall, the results showed that models trained or prompted with very few or no examples achieved higher overlap with the GT than those trained directly on the dataset. Indeed, the ICL setting consistently outperformed the baseline, SFT, and DPO across all styles, with higher overlap scores and stronger human agreement. By contrast, the baseline and SFT lagged behind, with SFT offering no significant improvement over the baseline. DPO, in turn, presented an interesting contrast. Despite featuring the lowest ground truth overlap scores, it managed to earn noticeably higher human agreement than the baseline and SFT. This suggests that its outputs, though less aligned with the target responses, exhibited more distinct stylistic marking than those generated by SFT. However, DPO still scored lower than ICL in both ground truth overlap and human agreement. In addition, the clarity subset stood out as the most challenging style, possibly due to less consistent stylistic cues, which made model outputs, and thus human judgments, more variable.

5.2 Limitations

Despite its contributions, this study has some limitations worth considering. First, the four communication styles explored (clarity, friendliness, empathy, and politeness) often overlap in both theory and practice. This overlap made it challenging to clearly separate them in the dataset, which may have influenced both the quality of the generated responses and the accuracy of the evaluation. Second, the use of open-ended prompts gave the model freedom to interpret each style in its own way. While this encouraged diversity, it also introduced ambiguity, making the stylistic distinctions less consistent. The lack of explicit style definitions may have led the model to conflate the different styles. Third, the dataset was relatively small, with 4,000 examples, and each response was limited to 30 words. This helped reduce computational cost, but also restricted the complexity and natural flow of the responses. Fourth, all annotations were done by a single rater due to time and resource constraints. Although ground truth agreement was measured, the absence of a second annotator means that potential bias was not controlled for, especially in the case of clarity, where agreement was particularly low (0.52), suggesting that even humans struggle to assess this dimension consistently. Fifth, due to time restrictions, only a small part of the linguistic data profiling (BERTScore and ROUGE-L) was used to evaluate the predictions of the text generation task. Finally, human-authored CS data were unavailable, preventing a direct comparison with the model-generated responses. As a result, any conclusions about how closely LLMs mirror real-world human communication styles should be interpreted with caution, as they are based solely on behaviors reported in prior literature.

5.3 Future Work

Future research could build on this work in several ways. One important step would be to collect longer, multi-turn synthetic conversations, which could better capture the dynamics of real CS interactions. Expanding the annotation process to include multiple and diverse annotators would also help improve the reliability and generalizability of the findings. Providing clearer and more detailed definitions for each style might reduce style overlap and make stylistic differences easier to identify. Another promising direction would be to directly compare synthetic and human-generated responses using the same annotation framework and evaluation tools. In addition, future studies could benefit from more advanced metrics and finer-grained classifiers, for example those targeting discourse structure or pragmatic intent, to better capture subtle stylistic features. Furthermore, the entire linguistic data profiling introduced in Section 3.2 could be used to evaluate the performance of stylistically marked text generation. This could yield interesting results and provide a clearer picture of the stylistic differentiation abilities of AI models. Lastly, it would be valuable to explore how language models handle stylistic variation in different languages and cultural contexts, where expectations and norms may differ significantly.

5.4 Final Remarks

Overall, this study demonstrates that LLMs are capable of generating CS responses that reflect stylistic nuance and many of the patterns discussed in prior research on human CS communication. At the same time, some communicative intents are realized in ways that, while not necessarily inappropriate or unnatural, diverge from what previous studies consider most effective in CS. This might be an indicator of the variability present in the human data these models are trained on. Moreover, the text generation experiment suggests that the dataset may contain noise or lack sufficient representativeness, as it led to greater semantic drift from the ground truth when it was used as training data than when only a few or no examples were provided. I believe that the findings discussed in the previous sections helped shed light on both the strengths and limitations of the LLMs' internal representations of style, as well as the sometimes unpredictable dynamics of synthetic text generation. These limitations should be carefully considered when deploying AI in CS-related applications, as they may lead to unintended chatbot responses, potentially lowering customer satisfaction and harming business outcomes. I also believe that these findings offer a useful starting point for future efforts to better align model outputs with the communicative values expected in real human interactions, both in CS and in broader contexts. It is my hope that this research supports continued work toward models that are not only effective in addressing user needs, but also capable of interpreting the pragmatic nuances of human interaction, ultimately enhancing user satisfaction and trust.

References

Babakov, N., Dale, D., Gusev, I., Krotova, I., & Panchenko, A. (2023). Don't lose the message while paraphrasing: A study on content preserving style transfer. In E. Métais, F. Meziane, V. Sugumaran, W. Manning, & S. Reiff-Marganiec (Eds.), *Natural language processing and information systems* (pp. 47–61). Springer Nature Switzerland.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Barcelos, R. H., Dantas, D. C., & Sénécal, S. (2018). Watch your tone: How a brand's tone of voice on social media influences consumer responses. *Journal of Interactive Marketing*, *41*(1), 60–80.

Braga, M., Kasela, P., Raganato, A., & Pasi, G. (2024). Synthetic data generation with Large Language Models for personalized community question answering. *arXiv preprint arXiv:2410.22182*.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list.

Busker, T., Choenni, S., & Bargh, M. S. (2025). Exploiting GPT for synthetic data generation: An empirical study. *Government Information Quarterly*, *42*(1), 101988.

Cameron, D. (1997). Language: Are you being served? *Critical Quarterly*, *39*(2).

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., & Jia, J. (2023). Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Cheng, X., Bao, Y., Zarifis, A., Gong, W., & Mou, J. (2021). Exploring consumers' response to text-based chatbots in e-commerce: The moderating role of task complexity and chatbot disclosure. *Internet Research*, *32*(2), 496–517.

Clark, C. M., Murfett, U. M., Rogers, P. S., & Ang, S. (2013). Is empathy effective for customer service? Evidence from call center interactions. *Journal of Business and Technical Communication*, *27*(2), 123–153.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505.

Creelman, V. (2022). "thank you for reaching out:" Brand relationship management and the conversational human voice of customer care in online service encounters. *Discourse, Context & Media*, *46*, 100572.

Fiscozen. (n.d.). Chi siamo. Retrieved March 24, 2025, from https://www.fiscozen.it/chisiamo/

Flesch, R. (1979). How to write plain English. *University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016]*.

Friginal, E. (2024). The case of task-oriented, polite discourse in intercultural aviation and customer service interactions. *Journal of Corpora and Discourse Studies*, *7*, 258–281.

Fuoli, M., Clarke, I., Wiegand, V., Ziezold, H., & Mahlberg, M. (2021). Responding effectively to customer feedback on Twitter: A mixed methods study of webcare styles. *Applied linguistics*, *42*(3), 569–595.

Gung, J., Moeng, E., Rose, W., Gupta, A., Zhang, Y., & Mansour, S. (2023). NatCS: Eliciting natural customer support dialogues. *arXiv preprint arXiv:2305.03007*.

Han, E., Yin, D., & Zhang, H. (2022). Chatbot empathy in customer service: When it works and when it backfires.

Hartmann, J. (2022). Emotion English DistilRoBERTa-base.

Hu, Y., Tafti, A., & Gal, D. (2019). Read this, please? The role of politeness in customer service engagement on social media.

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

Irvine, J. T. (2001). *"Style" as distinctiveness: The culture and ideology of linguistic differentiation*. na.

Jo, S., & Seo, J. (2024). ProxyLLM: LLM-driven framework for customer support through text-style transfer. *arXiv preprint arXiv:2412.09916*.

Kelleher, T. (2009). Conversational voice, communicated commitment, and public relations outcomes in interactive online communication. *Journal of communication*, *59*(1), 172–188.

Kiesling, S. F., & Schilling-Estes, N. (1998). Language style as identity construction: A footing and framing approach.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148–161.

Lewis, B. R., & Mitchell, V. W. (1990). Defining and measuring the quality of customer service. *Marketing intelligence & planning*, *8*(6), 11–17.

Li, X., Tang, H., Chen, S., Wang, Z., Chen, R., & Abram, M. (2024). Why does in-context learning fail sometimes? Evaluating in-context learning on open and closed questions. *arXiv preprint arXiv:2407.02028*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, 74–81.

Luckmann, T. (2008). On social interaction and the communicative construction of personal identity, knowledge and reality. *Organization Studies*, *29*(2), 277–290.

Lutzky, U. (2024). "Doesn't really answer my question...": Exploring customer service interactions on twitter. *International Journal of Business Communication*, *61*(1), 92–114.

Maisto, A. (2025). Collaborative storytelling and LLM: A linguistic analysis of automatically-generated role-playing game sessions. *arXiv preprint arXiv:2503.20623*.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*(2), 381–392.

McIsaac, C. M., & Sepe, J. F. (1996). Improving the writing of accounting students: A cooperative venture. *Journal of Accounting Education*, *14*(4), 515–533.

Mendonça, J., Pereira, P., Menezes, M., Cabarrão, V., Farinha, A. C., Moniz, H., Carvalho, J. P., Lavie, A., & Trancoso, I. (2023). Dialogue quality and emotion annotations for customer support conversations. *arXiv preprint arXiv:2311.13910*.

Merriam-Webster. (n.d.-a). Clear. In *Merriam-webster.com dictionary*. Retrieved March 18, 2025, from https://www.merriam-webster.com/dictionary/clear

Merriam-Webster. (n.d.-b). Correct. In *Merriam-webster.com dictionary*. Retrieved March 24, 2025, from https://www.merriam-webster.com/dictionary/correct

Merriam-Webster. (n.d.-c). Empathy. In *Merriam-webster.com dictionary*. Retrieved March 18, 2025, from https://www.merriam-webster.com/dictionary/empathy

Merriam-Webster. (n.d.-d). Friendly. In *Merriam-webster.com dictionary*. Retrieved March 18, 2025, from https://www.merriam-webster.com/dictionary/friendly

Merriam-Webster. (n.d.-e). Polite. In *Merriam-webster.com dictionary*. Retrieved March 18, 2025, from https://www.merriam-webster.com/dictionary/polite

Merriam-Webster. (n.d.-f). Style. In *Merriam-webster.com dictionary*. Retrieved March 24, 2025, from https://www.merriam-webster.com/dictionary/style

Ngo, L. V., Nguyen, T. N. Q., Tran, N. T., & Paramita, W. (2020). It takes two to tango: The role of customer empathy and resources to improve the efficacy of frontline employee empathy. *Journal of Retailing and Consumer Services*, *56*, 102141.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *20*(2), 139–156.

Oxford Learner's Dictionaries. (n.d.). Style. In *Oxford Learner's Dictionaries*. Retrieved March 24, 2025, from https://www.oxfordlearnersdictionaries.com/definition/english/style_1?q=style

Packard, G., & Berger, J. (2021). How concrete language shapes customer satisfaction. *Journal of Consumer Research*, *47*(5), 787–806.

Pinto, D. (2011). Are Americans insincere? Interactional style and politeness in everyday America.

Piwek, P., Beun, R.-J., & Cremers, A. (2008). 'proximal' and 'distal' in language and cognition: Evidence from deictic demonstratives in Dutch. *Journal of pragmatics*, *40*(4), 694–718.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review*, *8*(4), 364–382.

Ruytenbeek, N., & Decock, S. (2024). Expressing and responding to customer (dis) satisfaction online: New insights from discourse and linguistic approaches. *International Journal of Business Communication*, *61*(1), 3–17.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*(1), 101.

Sufi, F. (2024). Addressing data scarcity in the medical domain: A GPT-based approach for synthetic data generation and feature extraction. *Information*, *15*(5), 264.

Swain, V. D., Zhong, Q., Parekh, J. R., Jeon, Y., Zimmerman, R., Czerwinski, M., Suh, J., Mishra, V., Saha, K., Hernandez, J., et al. (2024). AI on my shoulder: Supporting emotional labor in front-office roles with an LLM-based empathetic coworker. *arXiv preprint arXiv:2411.02408*.

Tan, A. H. T., Muskat, B., & Johns, R. (2019). The role of empathy in the service experience. *Journal of Service Theory and Practice*, *29*(2), 142–164.

Unsloth. (2024). Llama-3.2-1B-Instruct [Accessed: 2025-05-28].

Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, *23*(7), 443–452.

Wang, C., Liu, X., & Awadallah, A. H. (2023). Cost-effective hyperparameter optimization for large language model generation inference. *International Conference on Automated Machine Learning*, 21–1.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, *20*(3), 470–477.

Wieseke, J., Geigenmüller, A., & Kraus, F. (2012). On the role of empathy in customer-employee interactions. *Journal of service research*, *15*(3), 316–331.

Zabava Ford, W. S. (1999). Communication and customer service. *Annals of the International Communication Association*, *22*(1), 341–375.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Appendix

A Example Rows from the RESPONSible Service Dataset

request	response0	response1
I purchased a vacuum cleaner online which was listed as lightweight, but it's much heavier than expected. I need assistance with a return process or exchange, please.	Thank you for contacting us! We apologize for the inconvenience. To proceed with a return or exchange, please visit our returns page, or call our helpline for assistance.	We are sorry the vacuum cleaner did not meet expectations. To return or exchange it, kindly access our returns page online, or speak with our customer service line directly.
I purchased a laptop last week that isn't functioning properly; the battery drains quickly and the keyboard has several unresponsive keys. Can you help resolve this issue promptly?	Thank you for contacting us about your laptop. We apologize for the inconvenience. Please provide your purchase details so we can arrange for repair or replacement to address your concerns.	We are sorry to hear about the problems with your laptop. Please share your purchase information, and we will expedite repair or replacement services to resolve the battery and keyboard issues.

Table 29: Clear Dataset

request	response0	response1
Hello, I'm facing issues with my internet connection. The signal keeps dropping intermittently. Can you help me troubleshoot this problem? I'd appreciate some guidance on resolving it. Thank you!	Hi there! We're sorry to hear about your internet issues. Let's look into this together. Please check your router's connections and restart it. We'll provide further assistance if needed.	Hello! We're here to help with your connection troubles. Have you tried restarting your router? Let's check that first together. Please let us know how it goes. We'll assist you further!
Hi, I purchased a coffee maker last week, but it's not functioning. I'd like to know the steps for a return or refund, please. Thank you.	Hello! We're sorry for the inconvenience with your coffee maker. Please visit our returns page for steps on refunds. Let us know if you need more assistance.	Hi there! We're truly sorry your coffee maker isn't working. We'd love to help you with a refund. Please check our return page. Reach out anytime for more support!

Table 30: Friendly Dataset

request	response0	response1
Hello, I ordered a package last week. It was supposed to arrive yesterday, but I haven't received it yet. Can you please look into this for me?	We're sorry for the inconvenience. Your package is being prioritized, and you'll receive it soon. Thank you for your patience and understanding while we resolve this issue.	I'm truly sorry for any frustration caused. We're urgently investigating the delay and will have it resolved. Your patience and understanding mean so much as we work to fix this.
Hello, I've noticed an unexpected charge on my recent bill. Could you please help me understand what happened and how we can resolve this issue? Thank you in advance.	I'm sorry to hear about the unexpected charge. Let me look into this for you and find out what's going on. We'll resolve this issue together promptly. Thank you.	I understand your concerns about the unexpected charge, and I'm truly sorry for any stress this has caused. Let's work together to quickly resolve the issue and ensure your peace of mind.

Table 31: Empathetic Dataset

request	response0	response1
Hello. I recently submitted feedback regarding your product. I would like to know when I can expect a reply or any further information regarding my feedback.	Dear customer, thank you for your feedback. We are currently reviewing it and will respond shortly. Your feedback is important to us. Thank you for your patience and understanding.	Dear valued customer, we truly appreciate your feedback and are giving it our utmost attention. We will get back to you with a detailed response soon. Thank you for your kind patience.
Hello, I recently ordered an item from your store, but I've not received it yet. Can you please help me with tracking this order? Thank you.	Dear Customer, thank you for reaching out. Please provide your order number, and we'll gladly assist you with tracking its shipment status promptly. Thank you for your patience.	Thank you for contacting us, valued customer. Kindly share your order number so we may expedite tracking information for you promptly. We appreciate your patience and understanding. Thank you sincerely.

Table 32: Polite Dataset

B List of Cohesive Devices Used in the Discourse Analysis

And, also, too, as well, in addition, furthermore, moreover, besides, what's more, for example, for instance, such as, like, including, namely, in particular, similarly, likewise, in the same way, equally, just as, in other words, that is to say, i.e., to put it another way, because, since, as, due to, owing to, for this reason, so, therefore, thus, consequently, as a result, hence, accordingly, in order to, so that, for the purpose of, with the aim of, but, however, although, even though, whereas, while, on the other hand, in contrast, yet, nonetheless, nevertheless, still, admittedly, of course, even so, while it is true that, or, alternatively, on the one hand, on the other hand, either, or, neither, nor, then, next, after that, subsequently, eventually, finally, at last, while, as, at the same time, meanwhile, during, before, previously, earlier, up to that point, until then, after, later, afterwards, since then, now, at present, currently, at that moment, by then, at that time, in conclusion, in summary, to sum up, overall, eventually.

C Prompts for the Text Generation Tasks

C.1 In-Context Learning (ICL)

```
Instruction: A customer service agent responded to a message from a customer in the following way.
Response: [GT-res0, train]
Rephrase the response to make it more [STYLE]. Only output the rephrased response. Do not add any explanation or extra
    ↪ text.
More [STYLE] response: [GT-res1, train]

...

Instruction: A customer service agent responded to a message from a customer in the following way.
Response: [GT-res0, test]
Rephrase the response to make it more [STYLE]. Only output the rephrased response. Do not add any explanation or extra
    ↪ text.
More [STYLE] response:
```

C.2 Supervised Fine-Tuning (SFT)

```
System: Rephrase the following customer service message to make it more [STYLE].
Only output the rephrased message. Do not add any explanation or extra text.
User: [GT-res0, train]
Assistant: [GT-res1, train]
```

C.3 Direct Preference Optimization (DPO)

```
System: Rephrase the following customer service message to make it more [STYLE].
Only output the rephrased message. Do not add any explanation or extra text.
User: [SFT-res1, train]
Assistant: [GT-res1, train]
```

C.4 Baseline (Zero-shot Chat Template)

```
System: Rephrase the following customer service message to make it more [STYLE].
Only output the rephrased message. Do not add any explanation or extra text.
User: [GT-res0, test]
Assistant:
```