Assessment and Improvement of Customer Service Speech with Multiple Large Language Models

So Watanabe*, Chee Siang Leow*, Junichi Hoshino†, Takehito Utsuro†, and Hiromitsu Nishizaki*
* University of Yamanashi, Japan
E-mail: s_watanabe@alps-lab.org, {leow,hnishi}@yamanashi.ac.jp Tel/Fax: +81-552208361
† University of Tsukuba , Japan
E-mail: jhoshino@esys.tsukuba.ac.jp, utsuro@iit.tsukuba.ac.jp

***Abstract*—This paper introduces a framework using multiple large language models (LLMs) to assess and enhance the customer service interactions of a staff in service industry. Effective communication with customers is pivotal for better customer satisfaction. To enhance these skills, precise and constructive feedback is crucial for customer service staff. This study employs multiple LLMs within a round-table discussion framework, named "ReConcile" to evaluate and suggest improvements for customer service dialogues. Proposed method scores a customer service speech and suggests suitable response. An subjective experiment was conducted where human subjects compared the effectiveness of customer service speech assessments and response suggestions generated by both a single LLM and the ReConcile method. Results showed that the scores with ReConcile were closer to human senses compared to a single LLM which indicate the suggestions for improving the customer service staff's speech.**

I. Introduction

In the service industry, such as restaurants, face-to-face communication skills are important not only to provide quality service and increase customer satisfaction, but also to avoid problems [1], [2]. However, training to acquire customer service skills takes a lot of time and costs. Traditional customer service training methods include role-playing and service manuals, but these training methods cannot replicate actual customer interactions. To address this issue, various training methods have been proposed to improve customer service skills training. For example, Furuno et al. [3] developed a system that uses virtual reality (VR) to evaluate behaviors such as bowing accuracy. Similarly, Nishio et al. [4] proposed a system that provides feedback on the use of filler words and fluency during customer service speech. However, these systems primarily focus on specific behaviors and do not address the modification and evaluation of response content, which is an essential aspect of effective service communication. To fill this gap, Sano et al. [5] developed a spoken dialogue system for advanced scenario-based customer service training. Although this system could evaluate the appropriateness of predefined utterances in a given scenario, it lacked the flexibility to adapt to unpredictable conversations during actual customer service interactions.

In customer service, the ability to communicate effectively with customers based on their specific situations is essential, especially in terms of content accuracy and respect for the customer. In addition, when considering customer service in Japanese, it is necessary to take into account the unique characteristics of the Japanese language. In Japanese, "*keigo*" (honorific) is a special way of expressing respect for the other party. In Japanese society, where the use of *keigo* is a general rule, the language used by service personnel can significantly affect the evaluation of customer service communication. Especially in the service industry, the use of honorifics and customer service that considers the customer's perspective are essential for successful Japanese language communication with customers. Therefore, this paper proposes a method for appropriately modifying and presenting the content of customer service language uttered by customer service staffs.

In recent years, with the emergence of large language models (LLMs) and their effective use in various domains, research on LLMs has progressed, and prompt engineering has also developed to generate more accurate responses by designing instructions and sentences to be entered into LLMs. Wei et al. [6] introduced a Chain of Thought (CoT) approach that embeds a stepwise process in prompts to elicit better responses from LLMs. There are studies that apply the approach of CoT. For example, Zero-shot CoT (Kojima et al. [7]), CoT-SC (Wang et al. [8] and Fu et al. [9]) and Tree of Thought (Yao et al. [10] and Long [11]). Furthermore, Bsharat et al. [12] presented 26 prompting principles that are essential for improving the response accuracy of LLMs. However, no study has used LLMs to evaluate and modify the appropriateness of customer service speech content. It is also unclear whether LLMs can provide appropriate feedback on honorifics that are unique to Japanese.

We propose a method for analyzing the speech of customer service staffs and modifying it to more appropriate expressions using the ReConcile method [13], which uses a combination of multiple LLMs, specifically GPT-4[1], Claude2 [14], and Bard (PaLM2) [15]. In the original paper [13], the effectiveness of the method was demonstrated on tasks requiring common sense reasoning and mathematical reasoning, such as StrategyQA[16], ECQA[17], GSM8K[18], and AQuA[19]. In recent years, it has become clear that multiple LLMs can be used to solve various tasks more effectively [20]–[22]. This paper also aims to more appropriately modify and evaluate customer service speech through ReConcile, which is based

[1] https://openai.com/blog/chatgpt

on a council system of outputs from multiple LLMs. Since the modification and evaluation of customer service responses do not have a single definitive answer, diverse expressions may be appropriate depending on the customer's needs and the store's situation. Therefore, it is necessary to generate optimal responses tailored to each specific situation.

ReConcile consists of three phases. At each phase, different LLMs are used to generate and modify responses, and finally, the optimal response is derived. First, the customer service situation settings, customer utterances and actions, and the customer service staff's responses to them, including prompts, are entered into ReConcile, and the customer service staff's responses are modified to generate the optimal response (response modification part). Next, the situation setting, the customer's utterance, and the staff's original and modified responses are re-entered into ReConcile, and the original response is rated on a 10-point scale, along with the generation of explanatory text on which the rating is based (scoring part).

To verify the effectiveness of ReConcile, an evaluation experiment was conducted with 20 subjects, 13 with customer service experience and seven with no customer service experience, to compare the results of response modification and scoring by a single LLM with those of ReConcile.

Subjects evaluated the response modification results from each method and selected the most appropriate response, with ReConcile rated as the most appropriate response modification in 57% of all evaluations. This was significantly higher than the second place score of 20% for GPT-4.

For both scoring and explanation, ReConcile ranked first with 55%, a 21 point improvement over GPT-4's 34%. These results indicate that ReConcile is able to modify responses and generate scores and explanations more appropriate to the customer service context than a single LLM. Moreover, ReConcile's modified responses and scores are at or near the same level as those generated manually, suggesting its ability to provide appropriate feedback in customer service training. The contribution of this study is that the ReConcile method, which uses multiple LLMs, has the potential to automatically analyze and correct the speech of customer service staffs and provide opportunities to learn appropriate customer service expressions, including nuanced honorific expressions. Through the generation of examples, persuasive scoring, and explanations, multiple LLMs such as ReConcile can be expected to contribute to the improvement of customer service staffs' communication skills.

The test dialogues data used in this paper are publicly accessible on Github[2].

II. Proposed Method

A. ReConcile

First, the flow of ReConcile is described: ReConcile performs three phases, which take advantage of the strengths of different large language models (LLMs) to obtain more accurate answers.The comprehensive process is illustrated in the Fig. 1.

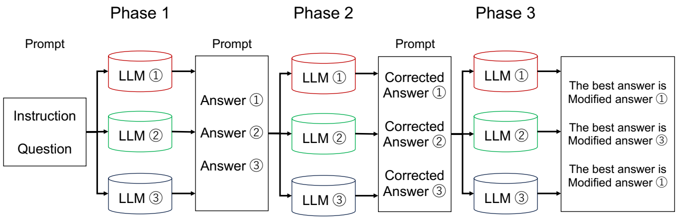


Fig. 1. A processing flow of ReConcile

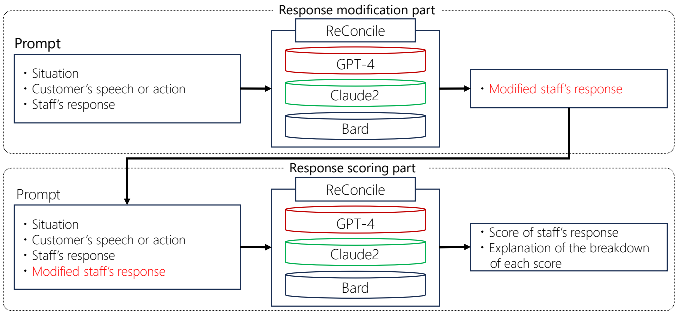

Response modification part
Prompt
・Situation
・Customer's speech or action
・Staff's response
ReConcile
GPT-4
Claude2
Bard
・Modified staff's response
Response scoring part
Prompt
・Situation
・Customer's speech or action
・Staff's response
・Modified staff's response
ReConcile
GPT-4
Claude2
Bard
・Score of staff's response
・Explanation of the breakdown of each score

Fig. 2. Flow of customer service evaluation using ReConcile

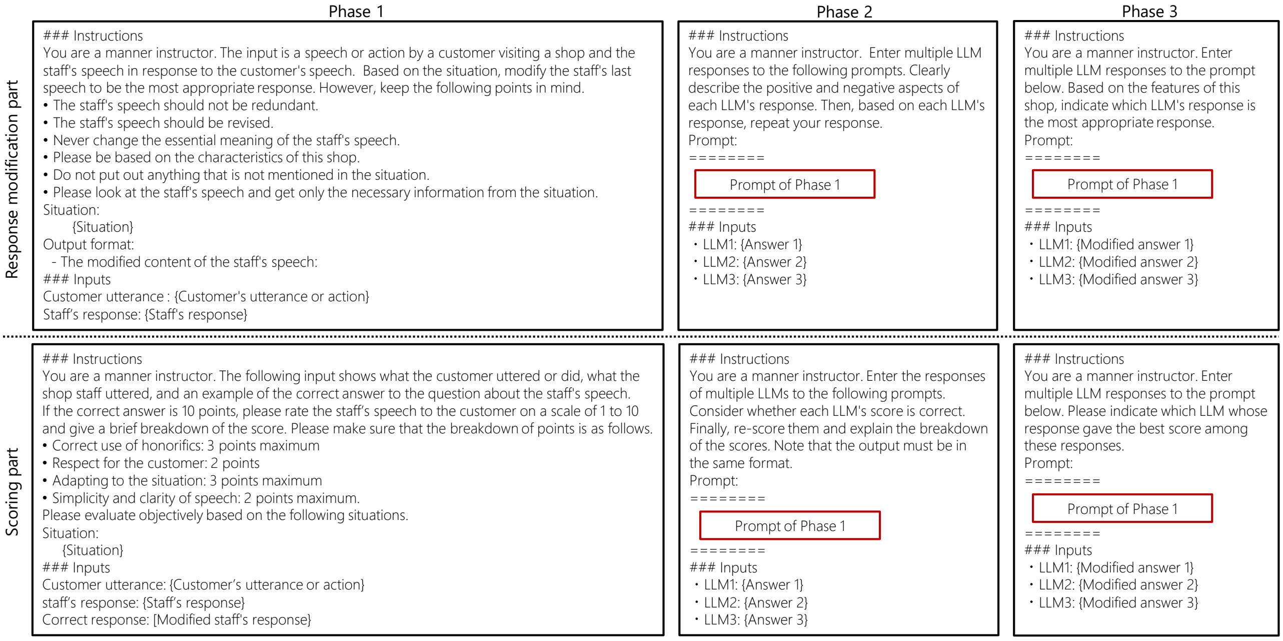
In Phase 1, a prompt consisting of instructions or queries is formulated and then submitted to multiple LLMs. Because LLMs are developed based on different data sets, their outputs embody unique perspectives and insights, even when responding to identical prompts. In Phase 2, the responses generated by each LLM in Phase 1 are merged into a new prompt, which is again input to each LLM. This process allows each LLM to output specific feedback for the multiple responses given, identifying the good and bad points of each response. In addition, each LLM generates responses again based on this feedback. In Phase 3, the responses modified in Phase 2 are compiled into a prompt and entered into each LLM. Each model is then tasked with selecting the most appropriate response from the revised options. If the optimal responses identified by each LLM differ, the Phase 2 responses are further evaluated and adjusted. Conversely, if the most appropriate responses identified by all LLMs converge in Phase 3, that response is considered the final answer. However, if consensus among all LLMs' responses remains elusive after up to five iterations of Phases 2 and 3, the final response is determined by majority vote.

B. Customer service evaluation using ReConcile

In this study, we use ReConcile to evaluate the content of staff's responses by modifying their statements to generate more appropriate responses. In addition, we score the customer service responses out of 10 and also generate explanations that serve as the basis for these scores. This entire process is illustrated in Fig. 2.

The prompts for each phase of ReConcile in both the response modification and scoring parts are shown in Fig. 3.

[2] https://github.com/watanabe0402/Customer-service-speech-assessment-with-ReConcile/tree/main



Instructions
You are a manner instructor. The input is a speech or action by a customer visiting a shop and the staff's speech in response to the customer's speech. Based on the situation, modify the staff's last speech to be the most appropriate response. However, keep the following points in mind.
• The staff's speech should not be redundant.
• The staff's speech should be revised.
• Never change the essential meaning of the staff's speech.
• Please be based on the characteristics of this shop.
• Do not put out anything that is not mentioned in the situation.
• Please look at the staff's speech and get only the necessary information from the situation.
Situation:
{Situation}
Output format:
- The modified content of the staff's speech:
Inputs
Customer utterance : {Customer's utterance or action}
Staff's response: {Staff's response}

Instructions
You are a manner instructor. Enter multiple LLM responses to the following prompts. Clearly describe the positive and negative aspects of each LLM's response. Then, based on each LLM's response, repeat your response.
Prompt:
========
Prompt of Phase 1
========
Inputs
・LLM1: {Answer 1}
・LLM2: {Answer 2}
・LLM3: {Answer 3}

Instructions
You are a manner instructor. Enter multiple LLM responses to the prompt below. Based on the features of this shop, indicate which LLM's response is the most appropriate response.
Prompt:
========
Prompt of Phase 1
========
Inputs
・LLM1: {Modified answer 1}
・LLM2: {Modified answer 2}
・LLM3: {Modified answer 3}

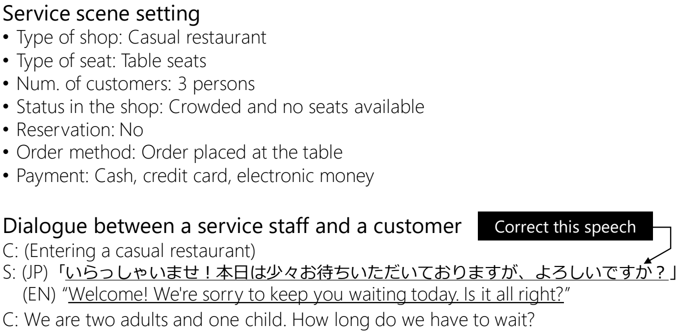
Instructions
You are a manner instructor. The following input shows what the customer uttered or did, what the shop staff uttered, and an example of the correct answer to the question about the staff's speech. If the correct answer is 10 points, please rate the staff's speech to the customer on a scale of 1 to 10 and give a brief breakdown of the score. Please make sure that the breakdown of points is as follows.
• Correct use of honorifics: 3 points maximum
• Respect for the customer: 2 points
• Adapting to the situation: 3 points maximum
• Simplicity and clarity of speech: 2 points maximum.
Please evaluate objectively based on the following situations.
Situation:
{Situation}
Inputs
Customer utterance: {Customer's utterance or action}
staff's response: {Staff's response}
Correct response: [Modified staff's response}

Instructions
You are a manner instructor. Enter the responses of multiple LLMs to the following prompts. Consider whether each LLM's score is correct. Finally, re-score them and explain the breakdown of the scores. Note that the output must be in the same format.
Prompt:
========
Prompt of Phase 1
========
Inputs
・LLM1: {Answer 1}
・LLM2: {Answer 2}
・LLM3: {Answer 3}

Instructions
You are a manner instructor. Enter multiple LLM responses to the prompt below. Please indicate which LLM whose response gave the best score among these responses.
Prompt:
========
Prompt of Phase 1
========
Inputs
・LLM1: {Modified answer 1}
・LLM2: {Modified answer 2}
・LLM3: {Modified answer 3}

Fig. 3. Prompts given at each phase in ReConcile. The prompts are presented in English, but in the experiment, the prompts were given in Japanese with the exact same meaning.

First, the store situation, the customer's utterance or action, and the staff's response to it are included in the prompt and entered into ReConcile. ReConcile modifies the staff's response to generate the most appropriate response. This is called the response modification part. Next, the store situation, the customer's utterance or action, the staff's original response, and the previously generated modified response are all included in the prompt and re-entered into ReConcile. ReConcile scores the staff's original response on a scale of 1 to 10. This scoring is based on four criteria: correct use of keigo (3-point scale), respect for the customer (2-point scale), adaptation to the situation (3-point scale), and simplicity and clarity of speech (2-point scale). In addition, it simultaneously generates an explanation as to why it received such a score. This is called the scoring part.



Service scene setting
• Type of shop: Casual restaurant
• Type of seat: Table seats
• Num. of customers: 3 persons
• Status in the shop: Crowded and no seats available
• Reservation: No
• Order method: Order placed at the table
• Payment: Cash, credit card, electronic money

Dialogue between a service staff and a customer
C: (Entering a casual restaurant)
S: (JP)「いらっしゃいませ！本日は少々お待ちいただいておりますが、よろしいですか？」
(EN) "Welcome! We're sorry to keep you waiting today. Is it all right?"
C: We are two adults and one child. How long do we have to wait?

Fig. 4. An example of a customer service scene setting and a customer service dialogue scenario. "C" is the customer's speech or action, and "S" is the service staff's speech. In reality, the LLMs modify the underlined Japanese speech.

III. Experiment

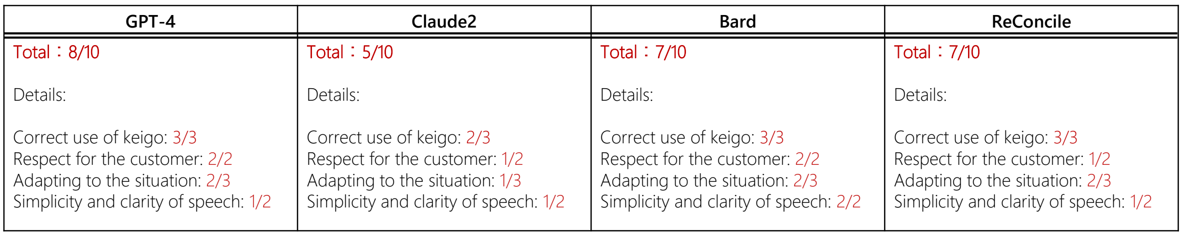
A. Experimental conditions

In the evaluation experiment, we assessed the effectiveness of ReConcile by comparing the responses from the individual GPT-4, Claude2, and Bard LLMs with those from the integrated ReConcile method. Furthermore, we evaluated the appropriateness of the response modified by ReConcile. The appropriateness evaluations were conducted through questionnaires administered to 20 subjects, 13 with customer service experience and seven without customer service experience.

1) Test dialogues data: A specific customer service scenario was created as evaluation data. The customer service scenarios were created by having ChatGPT play the role of the customer and a human play the role of the staff, interacting with the customer like real-life dialogues. In total, 12 customer service scenarios were created, including eight types of restaurants and four types of other stores. From the created scenarios, 30 dialogues that included characteristic interactions and complaint handling specific to the establishment were extracted, and an evaluation experiment was conducted using these dialogues.

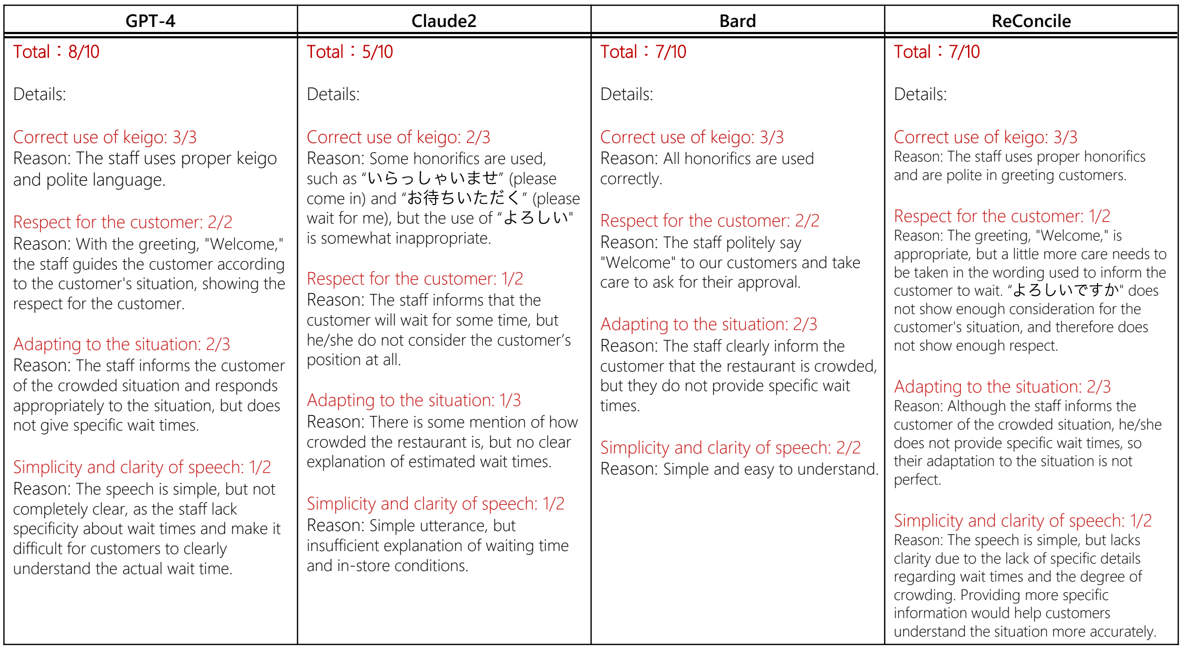
2) Ranking and Appropriateness Evaluation: The test dialogue data presented in Section III-A1 is shown in the underlined part of Fig. 4. Each LLM and ReConcile is used to make corrections to this staff's statements, assign a score, and generate an explanation of why the score is the way it is.

In order to evaluate the modifications proposed by each LLM (an example is shown in Table I) and the scores and their explanations shown in Fig. 6, subjects were asked four questions in a questionnaire format. First, in Question 1, the subjects were asked to rank the GPT-4, Claude2, Bard, and ReConcile modifications in terms of their suitability for customer service situations, accuracy of salutations, and clarity



GPT-4	Claude2	Bard	ReConcile
Total：8/10	Total：5/10	Total：7/10	Total：7/10
Details:	Details:	Details:	Details:
Correct use of keigo: 3/3 Respect for the customer: 2/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 1/2	Correct use of keigo: 2/3 Respect for the customer: 1/2 Adapting to the situation: 1/3 Simplicity and clarity of speech: 1/2	Correct use of keigo: 3/3 Respect for the customer: 2/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 2/2	Correct use of keigo: 3/3 Respect for the customer: 1/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 1/2

Fig. 5. Examples of scores of the customer service speech output from each LLM and the proposed method



GPT-4	Claude2	Bard	ReConcile
Total：8/10	Total：5/10	Total：7/10	Total：7/10
Details:	Details:	Details:	Details:
Correct use of keigo: 3/3 Reason: The staff uses proper keigo and polite language.	Correct use of keigo: 2/3 Reason: Some honorifics are used, such as "いらっしゃいませ" (please come in) and "お待ちいただく" (please wait for me), but the use of "よろしい" is somewhat inappropriate.	Correct use of keigo: 3/3 Reason: All honorifics are used correctly.	Correct use of keigo: 3/3 Reason: The staff uses proper honorifics and are polite in greeting customers.
Respect for the customer: 2/2 Reason: With the greeting, "Welcome," the staff guides the customer according to the customer's situation, showing the respect for the customer.	Respect for the customer: 1/2 Reason: The staff informs that the customer will wait for some time, but he/she do not consider the customer's position at all.	Respect for the customer: 2/2 Reason: The staff politely say "Welcome" to our customers and take care to ask for their approval.	Respect for the customer: 1/2 Reason: The greeting, "Welcome," is appropriate, but a little more care needs to be taken in the wording used to inform the customer to wait. "よろしいですか" does not show enough consideration for the customer's situation, and therefore does not show enough respect.
Adapting to the situation: 2/3 Reason: The staff informs the customer of the crowded situation and responds appropriately to the situation, but does not give specific wait times.	Adapting to the situation: 1/3 Reason: There is some mention of how crowded the restaurant is, but no clear explanation of estimated wait times.	Adapting to the situation: 2/3 Reason: The staff clearly inform the customer that the restaurant is crowded, but they do not provide specific wait times.	Adapting to the situation: 2/3 Reason: Although the staff informs the customer of the crowded situation, he/she does not provide specific wait times, so their adaptation to the situation is not perfect.
Simplicity and clarity of speech: 1/2 Reason: The speech is simple, but not completely clear, as the staff lack specificity about wait times and make it difficult for customers to clearly understand the actual wait time.	Simplicity and clarity of speech: 1/2 Reason: Simple utterance, but insufficient explanation of waiting time and in-store conditions.	Simplicity and clarity of speech: 2/2 Reason: Simple and easy to understand.	Simplicity and clarity of speech: 1/2 Reason: The speech is simple, but lacks clarity due to the lack of specific details regarding wait times and the degree of crowding. Providing more specific information would help customers understand the situation more accurately.

Fig. 6. Examples of scores and breakdown descriptions of the customer service speech output from each LLM and the proposed method

TABLE I
EXAMPLES OF STAFF'S CUSTOMER SERVICE SPEECH "WELCOME! WE'RE SORRY TO KEEP YOU WAITING TODAY. IS IT ALL RIGHT?" AS MODIFIED IN EACH LLM AND RECONCILE

	Modified staff's speech
GPT-4	Welcome! We are currently quite busy, so you will have to wait a while. Thank you for your understanding
Claude2	Welcome! We apologize for the inconvenience, but due to the current high volume of customers, there will be a short wait before we can seat you.
Bard	Welcome! Today, we are exceptionally busy, and there may be a bit of a wait. Would that be alright with you?
ReConcile	Welcome! Today, we are experiencing significant congestion, and there may be an approximate **10-minute** wait before we can guide you to your seat. We appreciate your patience and understanding.

of information. This will help us determine which method produced the best suggested revisions. Next, in Question 2, subjects were presented with only the score for the staff member's speech content by each method and asked to rate the appropriateness of the score with two choices, as shown in Fig.5. In Question 3, as shown in Fig. 6, subjects were presented with both the scores from Question 2 and explanations for these scores, and were asked to rank the methods again. By comparing the rankings with and without the score explanations, we can determine whether the subjects were satisfied with the scores. The final question, Question 4, provides a binary scale of whether the proposed modifications of ReConcile, the scoring, and its explanation are appropriate. Question 4 allows us to indicate whether ReConcile is able to make appropriate modifications and correctly explain the rationale for appropriate scoring and scores.

Furthermore, we also analyzed how correctly LLMs could point out honorifics in Japanese. For example, in one dialogue example, while GPT-4 modified the staff's response from "分かりました" (I understand) to "了解しました" (roger), ReConcile further refined it to the more appropriate "かしこまりました" (certainly, sir/ma'am). The phrase "了解しました" is often used by a superior giving permission or acknowledgment to a subordinate, making it unsuitable for use when addressing a customer, who should be treated with the utmost respect. On the other hand, "かしこまりました" is a more formal and polite expression in Japanese, commonly used in customer service to show deference and respect to the customer. Thus, it was shown that ReConcile can be modified to use honorifics more suitable for customer service.

In conclusion, the results of this study demonstrate that ReConcile, a method utilizing multiple LLMs, is superior to

TABLE II
RESULTS OF SUBJECT EXPERIMENTS FOR QUESTIONS 1 THROUGH 3 IN EACH LLM AND RECONCILE. NUMBERS ARE RATES [%]

	GPT-4	Claude2	Bard	ReConcile
[Q1] Rate ranked 1st (modified response)	20	5	17	57
[Q2] Appropriateness of scores (binary scale)	72	68	18	74
[Q3] Rate ranked 1st (scores and their breakdowns)	34	6	4	55

single LLM-based approaches in evaluating and improving customer service speech. ReConcile's ability to generate more appropriate modifications, accurate scores, and well-reasoned explanations, as well as its sensitivity to the nuances of Japanese honorifics, makes it a promising tool for improving customer service quality of customer service and communication skills.

B. Results

Table II shows the results of Questions 1 through 3, and shows that ReConcile was selected as the most appropriate method for correcting staff response speeches (57% of total responses), 37 percentage points higher than GPT-4 (20%). This indicates that ReConcile, which uses multiple LLMs, could modify responses more appropriately than a single LLM. Comparing the results of the staff's response modification by each method (Fig.I), ReConcile's response clearly indicates a specific waiting time of "10-minute wait". While this "10" figure may be a hallucination, what is important regardless is that a quantitative waiting time was presented. The provision of a quantitative waiting time allows customers to decide whether to wait or go to another store. In customer service, providing such specific information properly manages customer expectations, demonstrates store transparency, and is related to building customer trust. Therefore, ReConcile's revised response was deemed most appropriate.

ReConcile's scores were rated as "appropriate" more often than those of a single LLM, with a 2-point lead over GPT-4 alone, suggesting that ReConcile's scores are more appropriate. In addition, ReConcile was chosen first for scoring and commenting (55% of total responses), a 21-point improvement over GPT-4 (34%), suggesting that ReConcile scores more appropriately and generates more persuasive commenting than a single LLM.

Table III presents the results for Question 4, showing that ReConcile's responses are highly appropriate for modification (81%), scoring (78%), and commenting (82%), all exceeding those of a single LLM. These results show that ReConcile adapts to store situations, provides appropriate response modifications, and generates satisfactory explanatory text with appropriate scoring. Furthermore, 74% of the respondents found ReConcile's scoring to be "appropriate" when judged by the score alone (Table II). However, when the score and comment were considered together (Table III), 78% of subjects found the ReConcile score "appropriate", an improvement of 4 percentage points. This finding confirms that the explanations generated by ReConcile are more convincing in justifying the given scores.

TABLE III
QUESTION 4: RECONCILE'S APPROPRIATENESS FOR RESPONSE SPEECHES [%]

Appropriateness of the modification	Appropriateness of scores	Adequacy of commentary
81	78	82

IV. CONCLUSIONS

This paper proposed the use of ReConcile as an approach to evaluate and improve customer service language using multiple LLMs. ReConcile modifies staff's responses to specific customer service situations, provides examples of appropriate responses, and generates scores and explanations for the original responses. Subjective experiments showed that ReConcile outperforms single LLM-based methods in terms of response modification, scoring, and explanation generation, confirming its effectiveness in evaluating and improving customer service speech. It was also shown that ReConcile can handle Japanese-specific honorifics.

In the future, we plan to incorporate ReConcile into a customer service training system as a tool for response modification, scoring, and commentary generation.

ACKNOWLEDGMENT

This research was supported by JSPS Grant-in-Aid JP23H00994.

REFERENCES

[1] R. Asriyani and I. W. Anggayana, "Mastering the language of service: English communication skills for food and beverage professionals," *Jurnal Manajemen Pelayanan Hotel*, vol. 7, no. 2, pp. 1127–1139, 2023.

[2] S. Muda and S. M. Rashid, "Customer satisfaction towards communication skills of the franchise restaurant frontliners," in *Breaking the Barriers, Inspiring Tomorrow*, ser. European Proceedings of Social and Behavioural Sciences, vol. 110, 2021, pp. 223–230. DOI: 10.15405/epsbs.2021.06.02.30.

[3] T. Furuno, S. Fujita, D. Wang, *et al.*, "Multimodal vr customer service training system using conversational customer actors," *IPSJ Journal*, vol. 63, no. 1, pp. 231–241, 2022.

[4] T. Nishio, S. Iida, Y. Sano, *et al.*, "Customer service training vr system that can train how to speak," in *The Special Interest Group Technical Reports of IPSJ, Vol.2021-CVIM-224*, 2021, pp. 1–4.

[5] Y. Sano, C. S. Leow, S. Iiday, *et al.*, "Spoken dialog training system for customer service improvement," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 403–408.

[6] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 24 824–24 837.

[7] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 22 199–22 213.

[8] X. Wang, J. Wei, D. Schuurmans, *et al.*, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[9] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," in *The Eleventh International Conference on Learning Representations*, 2023.

[10] S. Yao, D. Yu, J. Zhao, *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 11 809–11 822.

[11] J. Long, "Large language model guided tree-of-thought," *arXiv preprint, arXiv:2305.08291*, 2023.

[12] S. M. Bsharat, A. Myrzakhan, and Z. Shen, "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4," *arXiv preprint, arXiv:2312.16171*, 2024.

[13] J. C.-Y. Chen, S. Saha, and M. Bansal, "Reconcile: Round-table conference improves reasoning via consensus among diverse llms," *arXiv preprint, arXiv:2309.13007*, 2023.

[14] Anthropic, "Model Card and Evaluations for Claude Models," *Technical Report 2023*, 2023.

[15] R. Anil, A. M. Dai, O. Firat, *et al.*, "Palm 2 technical report," *arXiv preprint, arXiv:2305.10403*, 2023.

[16] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, *Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies*, 2021. arXiv: 2101.02235 [cs.CL].

[17] S. Aggarwal, D. Mandowara, V. Agrawal, D. Khandelwal, P. Singla, and D. Garg, "Explanations for CommonsenseQA: New Dataset and Models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Aug. 2021, pp. 3050–3065.

[18] K. Cobbe, V. Kosaraju, M. Bavarian, *et al.*, *Training verifiers to solve math word problems*, 2021. arXiv: 2110.14168 [cs.LG].

[19] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program induction by rationale generation: Learning to solve and explain algebraic word problems," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 158–167. DOI: 10.18653/v1/P17-1015.

[20] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint, arXiv:2305.14325*, 2023.

[21] T. Liang, Z. He, W. Jiao, *et al.*, "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint, arXiv:2305.19118*, 2023.

[22] J. Li, Q. Zhang, Y. Yu, Q. Fu, and D. Ye, "More agents is all you need," *arXiv preprint, arXiv:2402.05120*, 2024.