Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

Support style-oriented in-context learning example selection for empathetic response generation during customer support conversations

W L Yeung

Saint Francis University, Hong Kong SAR, China

Abstract. With prompt-based in-context learning (ICL), LLMs have shown promise in various tasks including dialogue generation. Their performance, however, depends on the choice of ICL examples and hence the design of example selection method has become a major concern. This project considers the selection of in-context examples for response generation during customer support conversations. We design an example selection method which focuses on two typical support styles, namely emotional and informational, for improving the empathic quality of response generation. Through an experiment, we evaluate the method's effectiveness automatically in terms of the Epitome empathy metrics. The results show that the method brings different qualitative improvements with each support style.

Keywords: Dialogue generation · Response generation · Large language model · In-context learning

1 Introduction

Customer support (CS), also known as after-sales support or service, is an essential business function for addressing customer needs and enhancing satisfaction. Research has highlighted the "people" aspects of after-sales service as important in determining service quality and customer satisfaction [26, 17]. In particular, customers in need for help are satisfied by not only tangible (instrumental) support such as compensation, but also intangible support such as empathy which addresses customers' emotions [10, 5].

By modelling customer support conversations collected from various sources such as Twitter, machine learning researchers have demonstrated various ways to generate empathetic responses to customer requests [29, 23]. Meanwhile, large language models (LLMs) with their in-context learning (ICL) capability have shown promise in tackling the response generation task effectively [3, 15, 12, 27]. However, [13] showed that the performance of LLMs with ICL in various tasks is dependent on the choice of in-context examples used. This includes the task of empathetic response generation [12].

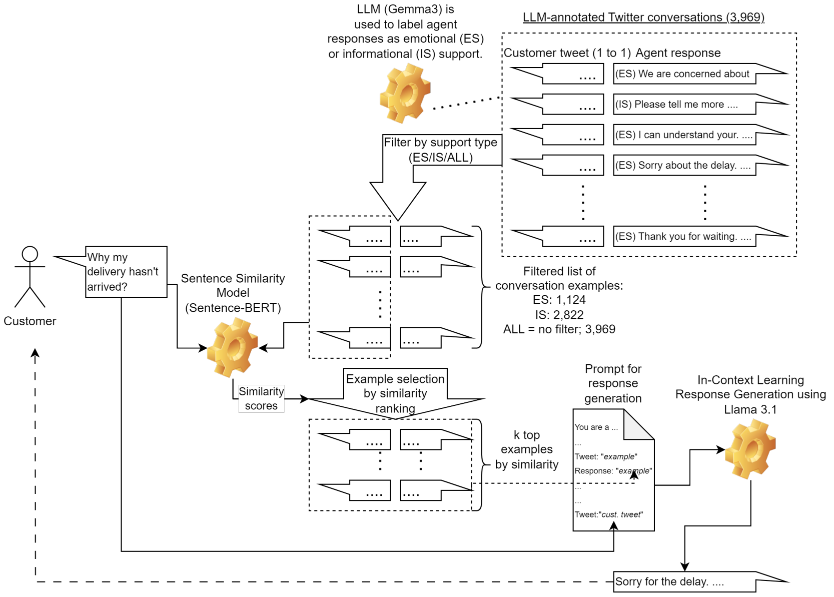
LLM (Gemma3) is used to label agent responses as emotional (ES) or informational (IS) support.
LLM-annotated Twitter conversations (3,969)
Customer tweet (1 to 1) Agent response
....
(ES) We are concerned about
(IS) Please tell me more ....
(ES) I can understand your. ....
(ES) Sorry about the delay. ....
(ES) Thank you for waiting. ....
Filter by support type (ES/IS/ALL)
Filtered list of conversation examples:
ES: 1,124
IS: 2,822
ALL = no filter; 3,969
Customer
Why my delivery hasn't arrived?
Sentence Similarity Model (Sentence-BERT)
Similarity scores
Example selection by similarity ranking
k top examples by similarity
Prompt for response generation
You are a ...
Tweet: "example"
Response: "example"
Tweet:"cust. tweet"
In-Context Learning Response Generation using Llama 3.1
Sorry for the delay. ....

Fig. 1. Response generation using prompt-based in-context learning with support style-oriented example selection

Following the lead of [12] in studying in-context example selection methods for LLM-based empathetic response generation, the present study experiments with a method of selecting response examples in the domain of customer support conversations. The selection method targets responses that convey emotional support and informational support as two main support styles and we evaluate the method's effects on the expression of empathy automatically based on the Epitome empathy metrics [21]. Figure 1 illustrates the overall response generation process.

The contributions of this work is as follows:

1. Design of a prompt-based ICL example selection method tailored for customer support response generation.
2. An automatic evaluation experiment on the method in terms of the empathic quality of the generated responses.

2 Related work

2.1 Emotional support and informational support

Empathy is often regarded as a useful ability or trait for frontline employees working in customer service call centres [4]. In particular, employees need empathy in order to show emotional support to customers in need for help [4].

In the context of customer support, [10] identified emotional support with things that are said and done to make the targets "*feel loved and bolster their sense of self-worth*" [10, p.25]. Apart from emotional support, they considered informational support as actual information that helps customers in need of assistance. Among various classifications of online support contents, emotional and informational support are deemed most commonly found [16, 8].

2.2 Response generation

Based on the deep learning approach to modelling conversations [25, 20, 22], researchers have demonstrated the generation of empathetic responses in customer service settings [2, 32, 29, 23]. More recently, researchers have looked into deploying pretrained language models through prompt-based in-context-learning (ICL) as a promising approach for the response generation task [31, 15, 12]. While ICL based on LLMs such as GPT-3 have been shown to perform competitively with fine-tuned language models in various tasks, [13] shows that their performance also depends on the choice of in-context examples. On the other hand, [12] focused on the task of empathetic response generation and studied the impact of example selection on GPT-3's performance in the task. Their results suggest an advantage of using few-shot (1 or 2 examples) learning over zero-shot learning.

2.3 Modelling and evaluating empathy

[21] demonstrated a computational approach to modelling empathy in conversations based on three communicative aspects of an empathetic response to an target person's utterance, namely:

Interpretation (IP) Show understanding of the target's feelings and experiences as inferred from the utterance.

Exploration (EX) Explore feelings and experiences not stated in the utterance.

Emotional reaction (ER) Express emotions such as warmth, compassion, and concern.

Furthermore, a response is rated as either strong, weak or missing in each of the three aspects. This modelling framework, known as Epitome, has been applied to studies in a number of domains including mental health support [21], distress support [28], counselling [14], dermatology support forum [11], and open-domain conversations [12, 1].

3 Method

We present in this section the design of our proposed support style-oriented example selection method for prompt-based in-context learning in a customer support response generation task. Details of an evaluation experiment on the selection method are also presented. Figure 1 illustrates the overall response generation process in the experiment.

```
You are reading tweets and replies from a customer service interaction.
Your task is to determine whether the last reply in the conversation
is offering emotional support or informational support based on the
content of the last reply.
If the last reply is comforting the customer, showing concern, making
them feel cared for or understood, or assuring the customer's self-worth,
respond with 'Emotional'.
If the last reply is providing information, advice, warnings, instructions,
or asking questions, respond with 'Informational'.
Otherwise, respond with 'Irrelevant'.

Conversation context:

 Customer tweet: "tweet"
 Agent reply: "reply"
```

Fig. 2. Prompt design for identifying emotional/informational support sentences using Gemma 3

3.1 Datasets and preprocessing

Following [30], the current study is based on a pre-existing Twitter customer support conversation dataset [6] which consists of over 10K conversations. We are interested in offering emotional support in a response's opening sentence and hence we focus on single-turn conversations (a single customer tweet followed by a single employee response), which amount to over 80% of the dataset. Furthermore, we tried to avoid short generic response sentences such as "Hello.", "Hi, [name].", "Glad to help." by arbitrarily excluding conversations with less than 4 words in the agent's opening sentence. We also ignored those conversations with a hyperlink embedded in the agent's opening response sentence.

The above preprocessing and filtering was applied to the original training and test sub-datasets with 10,000 and 500 conversations, respectively. Any test conversations that also appeared in the training data were removed. The results included 3,969 training and 181 test conversations. All resulting training conversations were employed as candidates for example selection and in-context learning during response generation.

3.2 Identifying emotional and informational support sentences

The Gemma 3 LLM [24] was leveraged for the task of identifying emotional support and informational support response sentences in the training dataset. We adapted the prompt for emotional classification from [9] for our task. Figure 2 shows our prompt design. Among the 3,969 agent responses in the training data, 1,124 were classified as emotional, 2,822 as informational support and 23 as irrelevant. Note that informational support sentences tend to be longer on average than emotional support sentences in the dataset. The average number of words in these responses are 8.77, 11.39 and 7.44 for emotional, informational and irrelevant support, respectively.

```
You are a professional customer service agent.
You are engaging in a conversation with a customer.
Respond in an empathetic manner using on average
11 words and maximum 37 words in one sentence
based on the following examples:

Tweet: “tweet”        { One-shot                  { Few-shot
Response: “reply”     { in-context learning       { in-context
_____________________ { (end of prompt)           { learning
⋮                                                 { (continuing)
Tweet: “tweet”                                    { (continuing)
Response: “reply”                                 { (end of prompt)
```

Fig. 3. Template for one-/few-shot prompt-based in-context learning response generation using Llama 3.1. For zero-shot generation, the template was shortened to only the first four lines together with the customer tweet.

3.3 In-context learning example selection

We applied the Sentence-BERT model [18] for selecting relevant candidate conversations from the training dataset for in-context learning (ICL) during response generation. For each conversation from the testing dataset, one or more candidate conversations were selected based on their sentence similarity scores with respect to the customer tweet in the test case. The experiment ran with the number of candidates set as 1 (one-shot ICL), then 3 and 5 (few-shot ICL). Furthermore, in order to gauge the effect of shortlisting candidate conversations based on their response support type (i.e. emotional vs. informational), the whole experiment was run based on the whole training dataset (ALL) as well as its two labelled subsets, namely, emotional-support (ES) and information-support (IS).

3.4 Response generation

We followed the approach of [12] to generate responses to customer tweets using an LLM with prompt-based in-context learning. We employed the Llama 3.1 LLM [7] in our experiment. Figure 3 shows our prompt design for response generation in the experiment.

For each test case, a total of 10 generated responses (together with the original human response) were included in the evaluation. The 10 modes of generation are:

- ZERO: zero-shot LLM prompt-based response generation (no examples provided).
- ALL1, ALL3, ALL5: LLM prompt-based response generation with 1-shot, 3-shot and 5-shot ICL based on *all* responses in the training dataset.

- ES1, ES3, ES5: LLM prompt-based response generation with 1-shot, 3-shot and 5-shot ICL based on *emotional-support* candidate responses in the training dataset.
- IS1, IS3, IS5: LLM prompt-based response generation with 1-shot, 3-shot and 5-shot ICL based on *information-support* candidate responses in the training dataset.

3.5 Evaluation metrics

We followed [12] and applied the Epitome and diversity metrics in our evaluation. The diversity of generated responses is based on the ratio of unique n-grams [19]. We also adapted the code[1] from [12] for conducting the evaluation.

4 Results and discussion

Table 1. Performance of response generation using different example selection methods

	Epitome			Diversity		Response Length	
	IP	EX	ER	dist-1	dist-2	Av. Len	Max Len
Human	0.000	0.055	0.801	0.975	1.000	11.796	35
ZERO	0.000	0.055	1.448	0.975	1.000	17.287	31
ALL1	0.000	0.287	1.448	0.972	1.000	17.094	31
ALL3	0.000	0.232	1.420	0.975	1.000	17.315	35
ALL5	0.000	0.265	1.409	0.970	1.000	17.724	33
ES1	0.000	0.298	1.387	0.980	1.000	15.276	31
ES3	0.000	0.232	**1.536**	0.976	1.000	16.122	28
ES5	0.000	0.210	**1.597**	0.975	1.000	16.652	30
IS1	0.000	**0.354**	1.470	0.975	1.000	17.132	35
IS3	0.000	**0.364**	1.370	*0.969*	1.000	**18.149**	31
IS5	0.000	**0.442**	1.343	*0.967*	1.000	**18.696**	38

(Numbers in **bold** are the highest in their respective columns.
Numbers in *italics* are the lowest in their respective columns.)

Table 1 summarises the performance of response generation in 10 different modes (and human performance) based on 181 test cases. First, all interpretation (IP) scores are nil in all 10 modes, as well as human responses. Second, with exploration (EX) scores, all non-zero-shot in-context learning (ICL) modes performed overwhelmingly better than the zero-shot ICL mode (ZERO) as well as the human responses. The selection of informational support examples for one-shot/few-shot in-context learning (IS1,IS3,IS5) achieved the highest EX scores.

On the other hand, according to the emotional reactions (ER) scores, all prompt-based ICL modes *including* the zero-shot (ZERO) mode performed better than the human responses. The selection of emotional-support style examples

[1] `https://github.com/passing2961/EmpGPT-3`

seemed to have some advantage over other modes but only in the few-shot cases (ES3 & ES5) and not the one-shot case (ES1).

On the whole, support style-oriented in-context examples contributed more to the explorative quality than to the emotional quality of prompt-based empathic response generation. The Llama LLM performed favourably against human responses in terms of emotional quality overall and was only boasted by few-shot (but not one-shot) learning with emotional-support style examples. [12] also found some advantages of few-shot learning over zero-shot learning in the related study based on GPT3. Finally, given that IS examples are on average longer than ES examples (11.39 vs 8.77), few-shot learning with IS examples (IS3 & IS5) tended to generate the longest responses, but also the lowest 1-gram diversity.

Table 2. A sample of responses generated in different modes in the experiment

Mode	Generated Response	IP	EX	ER
ZERO	I'm so sorry that your meeting recording isn't appearing, let me personally assist you in locating it.	0	0	2
ALL1	I understand your concern, and I'm here to assist you in finding that recording; let's work together to locate it for you.	0	0	1
ALL3	I'm here to assist you. Can you please provide more details about your recording, such as the meeting title and date?	0	2	1
ALL5	I'm here to assist you! Can you please share more details about the meeting recording, such as the date and time it was recorded?	0	2	1
ES1	I'm here to assist you; please tell me more about this meeting recording.	0	0	1
ES3	I'm sorry to hear that your meeting recording isn't appearing; let's work together to locate it for you!	0	0	2
ES5	I'm sorry to hear that your recording isn't displaying; let's work together to locate it for you!	0	0	2
IS1	I understand your frustration, let me help you locate that recording from the cloud.	0	0	1
IS3	I'm here to assist you, let's search for your meeting recording together. Can you please provide more details about the meeting, like date and time?	0	2	1
IS5	I'm happy to assist with finding your recording. Can you please provide more details about the meeting and the cloud storage service used?	0	2	1

Table 2 shows a sample of responses generated with the different selection methods together with their Epitome scores. As the sample shows, we can see more use of emotion words such as "sorry" in responses based on emotional support (see ES3 & ES5) whereas those based on informational support tend to involve asking questions (see IS3 & IS5).

5 Conclusion and further work

This research proposes a method of selecting examples for in-context learning during prompt-based empathetic response generation in customer support conversations. We have experimented with the method using a Twitter dataset and conducted automatic evaluation on it. The preliminary results suggest that while the LLM-based generation approach performs favourably against human responses in terms of emotional reaction, in-context learning with informational support-style examples can help render chatbots more empathetic with responses that explore customers' feelings and experiences.

Further work may include: (1) Supplementing the automatic evaluation with human evaluation in the experiment; (2) Increasing the number and size of Twitter dataset(s) in the experiment from other sources; (3) Including conversation data from other online platforms such as support forums; and (4) Experimenting with more recently released LLMs such as DeepSeek and GPT-OSS. Finally, in order to validate any empathy effects in real life, a further experiment may be conducted in the field (e.g., an actual customer service operation) to measure the effect of the proposed method on the effectiveness of LLM-based chatbots in affecting customer behaviour.

Acknowledgement The author would like to thank the anonymous reviewers for their comments and suggestions on an earlier version of this paper.

References

1. Amjad, B., Zeeshan, M., Beg, M.O.: Emp-eval: A framework for measuring empathy in open domain dialogues. arXiv preprint arXiv:2301.12510 (2023)
2. Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L.: Affective neural response generation. In: Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40. pp. 154–166. Springer (2018)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Clark, C.M., Murfett, U.M., Rogers, P.S., Ang, S.: Is empathy effective for customer service? Evidence from call center interactions. Journal of Business and Technical Communication **27**(2), 123–153 (2013)
5. Fuoli, M., Clarke, I., Wiegand, V., Ziezold, H., Mahlberg, M.: Responding effectively to customer feedback on twitter: a mixed methods study of webcare styles. Applied Linguistics **42**(3), 569–595 (2021)
6. Ganhotra, J., Roitman, H., Cohen, D., Mills, N., Gunasekara, C., Mass, Y., Joshi, S., Lastras, L., Konopnicki, D.: Conversational document prediction to assist customer care agents. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 349–356 (2020)
7. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

8. Huang, K.Y., Nambisan, P., Uzuner, O.: Informational support or emotional support: preliminary study of an automated approach to analyze online support community contents. In: Proceedings of 2010 International Conference on Information Systems (ICIS) (2010)
9. Imran, M.M., Chatterjee, P., Damevski, K.: Uncovering the causes of emotions in software developer communication using zero-shot llms. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. pp. 1–13 (2024)
10. Knight, M., Carpenter, S.: Optimal matching model of social support: An examination of how national product and service companies use twitter to respond to consumers. Southwestern Mass Communication Journal **27**(2) (2012)
11. Lee, G., Parde, N.: Acnempathize: A dataset for understanding empathy in dermatology conversations. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 143–153 (2024)
12. Lee, Y.J., Lim, C.G., Choi, H.J.: Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 669–683 (2022)
13. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., Chen, W.: What makes good in-context examples for GPT-3? In: Agirre, E., Apidianaki, M., Vulić, I. (eds.) Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. pp. 100–114. Association for Computational Linguistics, Dublin, Ireland and Online (May 2022). https://doi.org/10.18653/v1/2022.deelio-1.10, https://aclanthology.org/2022.deelio-1.10/
14. Lozoya, D.C., Lúa, E.H., Perches, J.A.B., Conway, M., D'Alfonso, S.: Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions. In: Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025). pp. 157–171 (2025)
15. Madotto, A., Lin, Z., Winata, G.I., Fung, P.: Few-shot bot: Prompt-based learning for dialogue systems. arXiv preprint arXiv:2110.08118 (2021)
16. Pfeil, U.: online support communities. Social Computing and Virtual Communities p. 121 (2009)
17. Rafaeli, A., Ziklik, L., Doucet, L.: The impact of call center employees' customer orientation behaviors on service quality. Journal of Service Research **10**(3), 239–255 (2008)
18. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
19. See, A., Pappu, A., Saxena, R., Yerukola, A., Manning, C.D.: Do massively pretrained language models make better storytellers? In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 843–861 (2019)
20. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1577–1586 (2015)

21. Sharma, A., Miner, A., Atkins, D., Althoff, T.: A computational approach to understanding empathy expressed in text-based mental health support. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5263–5276 (2020)
22. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, W.B.: A neural network approach to context-sensitive generation of conversational responses. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 196–205 (2015)
23. Spring, T., Casas, J., Daher, K., Mugellini, E., Abou Khaled, O.: Empathic response generation in chatbots. In: Proceedings of 4th Swiss Text Analytics Conference (SwissText 2019), 18-19 June 2019, Wintherthur, Switzerland. No. CONFERENCE, 18-19 June 2019 (2019)
24. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
25. Vinyals, O., Le, Q.V.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
26. Vredenburg, H., Wee, C.H.: The role of customer service in determining customer satisfaction. Journal of the Academy of Marketing Science **14**, 17–26 (1986)
27. Welivita, A., Pu, P.: Are large language models more empathetic than humans? arXiv preprint arXiv:2406.05063 (2024)
28. Welivita, A., Yeh, C.H., Pu, P.: Empathetic response generation for distress support. In: Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 632–644 (2023)
29. Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R.: A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 3506–3510 (2017)
30. Yeung, W.L.: Selecting empathic response headers in customer support conversations with llm-based emotion recognition. In: International Symposium on Chatbots and Human-Centered AI. pp. 23–32. Springer (2024)
31. Zheng, C., Huang, M.: Exploring prompt-based few-shot learning for grounded dialog generation. arXiv preprint arXiv:2109.06513 (2021)
32. Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)