Fine-tuning Open-source Large Language Models for Automated Response to Customer Feedback

M. Albuquerque[1], L. Barbosa[1], J. Moreira[1], A. da Silva[2], T. Melo[3]

[1] Universidade Federal de Pernambuco, Brazil
mvca@cin.ufpe.br, luciano@cin.ufpe.br, jms5@cin.ufpe.br
[2] Universidade Federal do Amazonas, Brazil
alti@icomp.ufam.edu.br
[3] Universidade Estadual do Amazonas, Brazil
tmelo@uea.edu.br

Abstract. Online reviews play a key role in influence customer decisions during their purchase journey. Consequently, negative feedback from customers can have an adverse impact on the sales of products or services, potentially leading to diminished revenue and market share. However, this effect can be mitigated by crafting thoughtful responses to these comments. This paper proposes using open-source pre-trained large language models, specifically smaller versions, to respond to negative reviews effectively. These models, pre-trained on large datasets, require minimal additional data for fine-tuning. To validate the effectiveness of this approach, we apply our solution to the domain of restaurant reviews. Our research shows that these fine-tuned models perform comparably to larger models, such as ChatGPT-3.5, in generating respectful, specific, and corrective responses that encourage customers to revisit the restaurant.

CCS Concepts: • **Computing methodologies** → **Machine learning**.

Keywords: large language model, natural language processing, supervised fine-tuning

1. INTRODUCTION

The rapid advancement of machine learning in NLP has led to increased interest in language models [Qiu et al. 2024; Clavié et al. 2023; Alnuhait et al. 2023]. Models like ChatGPT, a Generative Pre-trained Transformer (GPT) model, perform exceptionally well in tasks such as text summarization, translation, question answering, code generation, and customer support [Brown et al. 2020]. However, proprietary models like ChatGPT incur costs that might be prohibitive depending on the application's purpose and budget. Simultaneously, the development of robust pre-trained models, especially open-source ones, has grown [Zhao et al. 2023]. These models, pre-trained on vast datasets using unsupervised techniques, can be fine-tuned for specific tasks despite not being trained for all instructions [Devlin et al. 2019].

E-commerce has become a primary method for selling products and services, increasing online interactions and public reviews. Negative reviews significantly impact customer decisions more than positive ones [Lee et al. 2008]. Addressing negative reviews can improve purchase intent and trust [Qing et al. 2018; Sparks et al. 2016], especially when using conversational language and prompt responses. Studies show that fine-tuning pre-trained models effectively generates responses to customer reviews [Cao and Fard 2021]. Fine-tuning reduces the required data volume, leveraging the extensive unsupervised training these models undergo.

This article proposes using large language models to respond to customer feedback. We employ open-source language models and fine-tuning techniques like QLoRA [Dettmers et al. 2023] to re-

duce computational costs. Our methodology includes extracting customer comments from various platforms, processing them with ChatGPT to generate responses, and fine-tuning pre-selected open-source models. We evaluate these models using BLEU, ROUGE, and human evaluation. The primary contribution is demonstrating that targeted fine-tuning with a limited dataset enables smaller-scale models to perform comparably to larger ones for specific tasks.

The remainder of the article is organized as follows. Section 2 provides a review of the related work on Large language models (LLMs) and Section 3 presents an overview of the methodology applied in our study. Section 4 includes experimental evaluation of the proposed approach. Finally, Section 5 discusses our main conclusions, limitations, and future research directions.

2. RELATED WORK

2.1 Applications of LLMs in Customer Service

Large Language Models (LLMs), like GPT-3, are widely used to automate customer service responses [Wang et al. 2024; Schwartz et al. 2023]. These models generate contextually relevant and syntactically coherent responses due to their extensive training on diverse data sets. Studies show their effectiveness in chatbots and virtual assistants on various platforms [Zhang et al. 2019; Liu et al. 2023]. The use of LLMs in customer service streamlines interactions and ensures consistent communication, crucial for customer satisfaction and trust. Additionally, these models can be customized for specific business needs and languages, enhancing their global applicability [Clavié et al. 2023]. LLMs also automate and personalize customer interactions across sectors. Cao and Fard (2022) [Cao and Fard 2021] explored using pre-trained neural language models for automatic responses in mobile apps, improving user engagement and efficiency. These models' ability to understand and process natural language has been utilized in systems handling a wide range of customer service scenarios, from complaint resolution to product support [Sadiq et al. 2024]. Moreover, the evolution of LLMs has enabled multi-lingual capabilities, enhancing accessibility and inclusivity in global customer service solutions [Ahuja et al. 2023].

2.2 Sentiment Analysis and Response Generation

Sentiment analysis is vital in automating customer service by detecting customer sentiments to generate tailored responses. This technique classifies text into sentiments like positive, negative, or neutral, offering insights into customer emotions and opinions [Lee et al. 2008]. LLMs show significant potential in interpreting nuanced sentiments in feedback [Qing et al. 2018]. These insights help generate context-appropriate, empathetic, and personalized responses. For instance, Zhang et al. (2022) [Zhang et al. 2023] introduced a Transformer-based model for user reviews in mobile apps, integrating ratings and comments to generate accurate and relevant responses, enhancing customer satisfaction and engagement. Integrating sentiment analysis with response generation faces challenges, such as ensuring relevance and appropriateness, especially with complex emotions or complaints. Studies like [Gao et al. 2021] have explored advanced methodologies that enable LLMs to generate more context-aware and situation-specific responses, improving the effectiveness of automated systems in real-world customer service settings.

3. METHODOLOGY AND DEVELOPMENT

In this section, we detail each stage of the methodology employed in our research, including the critical decisions made throughout the development process. The stages outlined in the flowchart in Figure 1 are as follows:

(1) Data Extraction and Cleaning;

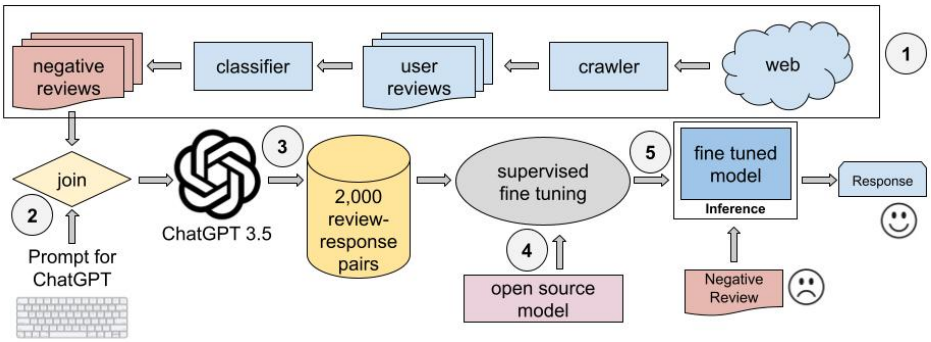


Fig. 1: Methodology overview.

(2) ChatGPT Prompt Development;
(3) Fine-tuning data generation using ChatGPT;
(4) Selection of pre-trained open-source models;
(5) Models fine-tuning.

3.1 Data Extraction and Cleaning

Customer reviews were collected from various platforms including Instagram, Facebook, iFood, Google Reviews, and TripAdvisor, covering 2012 to mid-2023. Several Brazilian restaurants were selected for this study, using web scraping techniques with BeautifulSoup [Richardson 2007] and Apify[1]. After data extraction, a cleaning process removed empty, noisy, or non-Portuguese comments, ensuring relevant information for analysis. This resulted in 2,000 negative customer reviews. We built a classifier to distinguish critical (negative) reviews from positive and neutral ones, necessary as some sources like Instagram and Facebook do not provide ratings. The classifier calculates the average polarity of terms using the lexicon cited in [de Melo 2022]. If the value is greater than zero, the comment is positive; otherwise, it is negative.

3.2 ChatGPT Prompt Development

After extracting the reviews, the next step is to develop a prompt for ChatGPT to generate clear, empathetic, and respectful responses to negative customer reviews. The aim is to address the customer's issues convincingly without being generic. To prevent customer reviews from altering the task, the comment is enclosed in < and > symbols, ensuring the prompt's structure and intent remain intact. Instructions ensure the response is empathetic, respectful, and not generic, leveraging ChatGPT's ability to generate text in various tones. The prompt clearly defines ChatGPT's role and objectives, guiding it to use the customer's review details effectively [Clavié et al. 2023]. The constructed prompt for ChatGPT is shown in Table I.

3.3 Fine-tuning data generation using ChatGPT

After constructing the prompt and incorporating customer feedback, we established instructions for ChatGPT using the ChatGPT-3.5 model via the OpenAI API. The generated responses were manually analyzed to ensure they met the input prompt criteria, serving as the ground truth for each client's feedback. We obtained a set of 2,000 customer review comments and responses, divided into training, validation, and test subsets with an 8:1:1 split ratio. This resulted in 1,600 pairs for training and 200 pairs each for validation and testing.

[1]https://apify.com

Table I: Prompt designed to generate responses with GPT-3.5, written in Portuguese.

```
Você é uma IA especializada em responder comentários negativos de um cliente a um
restaurante. Sua tarefa é responder respeitosamente um comentário negativo de
um cliente ao seu restaurante. Dado o comentário do cliente entre <>, gere um
comentário de resposta de forma respeitosa, empática e não genérica, convencendo
o cliente que medidas serão tomadas para resolver o seu problema e que ele poderá
voltar a fazer pedidos no restaurante. Certifique-se de usar detalhes específicos
do comentário do cliente.
<COMENTÁRIO>
```

Table II: Sample responses from pre-trained models.

Model	**Response**
Falcon 7B	*Obrigado por nos dar a oportunidade de melhorar nosso serviço.*
LLaMA 2 7B	*Olá, Como você pode ver, o cliente está muito desapontado com o atendimento que recebeu.*
Open-LLaMA 7B	*Agradecemos a sua opinião.*

3.4 Selection of pre-trained open-source models

From the literature review conducted during the article development, the three best open-source pre-trained models were selected based on their performance in HuggingFace benchmark tests. These tests assess the models' performance in answering elementary science questions, common-sense inference, multi-task scenarios with 57 tasks, and their tendency to reproduce fake news. The chosen models were Falcon, LLaMA 2, and OpenLLaMA, with versions ranging from 7 billion to 70 billion parameters. To minimize computational cost, the 7-billion-parameter versions were selected for fine-tuning. A preliminary analysis was conducted on these pretrained models' ability to generate responses to a negative customer review. A sample response for the review "*Alimentação caríssima para um péssimo atendimento.. super mal atendido*" can be seen in Table II. As shown in Table II, the models can identify some information related to the desired task. However, the generated responses are generic and not specific to the customer's complaint or issue. This highlights a key limitation of relying solely on pre-trained models without additional customization. While prompt engineering can guide models to perform specific tasks, it does not necessarily outperform fine-tuning [Shin et al. 2023]. Fine-tuning involves further training pre-trained models on customized data, allowing them to adapt to specific requirements. This process updates the pre-trained parameters with task-specific data, resulting in more accurate and relevant responses. Therefore, to achieve concise and stable responses that directly address customer issues, we have chosen the fine-tuning approach.

3.5 Models fine-tuning

Each selected pretrained model underwent supervised fine-tuning to refine its capabilities for responding to customer reviews. QLoRA [Dettmers et al. 2023] was used, allowing fine-tuning with a 4-bit quantized model, reducing model size by four times. This enabled training on a single NVIDIA RTX 3080 GPU with 10GB VRAM. The QLoRA setup included a rank of 64, LoRA scaling factor of 16, and dropout probability of 0.1. LoRA adapters were applied to both Query and Value matrices. Fine-tuning used a learning rate of 2×10^{-4}, a warmup ratio of 0.03, AdamW optimization, and cross-entropy loss, with a batch size of 16. Tokenization used the pretrained model's tokenizer. Training was conducted over one epoch, using the validation subset to select the best model version based on minimized validation loss. The best validation loss for each model is shown in Table III. The Open LLaMA model had the lowest loss in both training and validation subsets. However, while this metric indicates quality, it does not fully determine the model's ability to generate responses to customer comments, as it measures how well the model predicts the next token, not the entire complexity of

Table III: Loss value of each model in the validation set.

Model	Falcon 7B	LLaMA 2 7B	Open LLaMA 7B
Best Loss in Validation Set	0.94	0.81	0.78

Table IV: Sample of Responses Generated after Fine-Tuning.

Customer Review	**Alimentação caríssima para um péssimo atendimento.. super mal atendido**
Falcon 7B	Caro cliente, lamentamos profundamente que sua experiência em nosso restaurante tenha sido insatisfatória devido ao atendimento e à qualidade da comida. Agradecemos por compartilhar sua opinião e gostaríamos de entender melhor os pontos negativos mencionados para melhorar nossos serviços.
LLaMA 2 7B	Caro cliente, lamentamos profundamente pela experiência negativa que você teve em nosso restaurante. Pedimos desculpas pelo atendimento insatisfatório e pela alta cobrança. Vamos revisar nossos processos internos para garantir que isso não aconteça novamente. Agradecemos seu feedback e esperamos ter a oportunidade de oferecer uma experiência melhor no futuro.
Open LLaMA 7B	Olá, agradecemos pelo seu feedback. Lamentamos muito pela experiência negativa que você teve em nosso restaurante. Vamos analisar o que aconteceu para garantir que isso não aconteça novamente. Agradecemos por nos informar e esperamos ter a oportunidade de recebê-la novamente para uma experiência melhor.

Table V: Metrics of Models on the Test Set.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Falcon 7B	51.10	37.82	29.56	23.85	40.26
LLaMA 2 7B	49.86	37.11	29.11	23.53	40.21
Open LLaMA 7B	48.79	35.67	27.73	22.30	38.80

text generation. After the fine-tuning process was completed, new inferences were made using each fine-tuned model to verify the response generation capabilities acquired during training. A sample of the responses generated can be seen in Table IV (in Portuguese).

4. EVALUATION AND COMPARISON OF THE FINE-TUNED MODELS

After fine-tuning the models, each one acquired the ability to properly respond to negative customer reviews. This raised a new question about how good the responses generated are. To address this question, two distinct methods are proposed for evaluating and comparing the models: **evaluation through reference metrics** and **human evaluation**. Evaluation through reference metrics involves the use of quantitative techniques to generate a comparative value between the response generated by the model evaluated and a reference response. In our case, the reference response, considered ideal, was generated by ChatGPT-3.5. Human evaluation requires assessment by human individuals to determine the quality of the responses generated by the models in various scenarios. The subsequent subsections discuss each of these approaches in detail.

4.1 Evaluation Through Reference Metrics

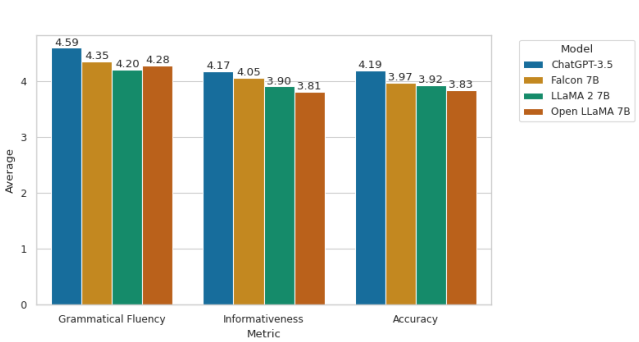
In this method, BLEU and ROUGE metrics quantify and compare models. BLEU, originally for evaluating text translations [Papineni et al. 2002], and ROUGE, designed for text summaries [Lin 2004], have proven useful for assessing responses to comments [Cao and Fard 2021; Zhang et al. 2023; Gao et al. 2021; Farooq et al. 2020]. The BLEU metric will be applied with n-grams ranging from 1 to 4, allowing analysis from single tokens to groups of four tokens in generated responses. For the ROUGE

metric, ROUGE-L will be used, calculating similarity by finding the longest common subsequence between reference and generated sentences, capturing complex semantic relationships. The test subset will be used to calculate these metrics, with ChatGPT-3.5's responses as the reference. Metric values, expressed in percentage, are presented in Table V. Overall, the models' performances on the metrics were similar. The Falcon model excelled, showing the best performance with BLEU across all n-gram variations and slightly outperforming LLaMA 2 on ROUGE-L. These results suggest that Falcon 7B effectively captures grammatical and content aspects close to the reference response, critical for applications requiring accurate content generation. Performance on BLEU-4, which considers longer n-grams, is notably lower than BLEU-1 for all models, indicating difficulty in maintaining coherence in longer text sequences. This suggests that while models can replicate short phrases well, they struggle with longer responses. It is important to note that these metrics, while useful for verifying a model's ability to generate responses, do not necessarily indicate its ability to generate specific responses to customer comments. A model may score high on these metrics but still produce generic responses, which is not ideal for the intended application [Cao and Fard 2021]. Given the limitations of these metrics, a Human Evaluation is proposed to better estimate the models' capabilities, which will be discussed in 4.2.

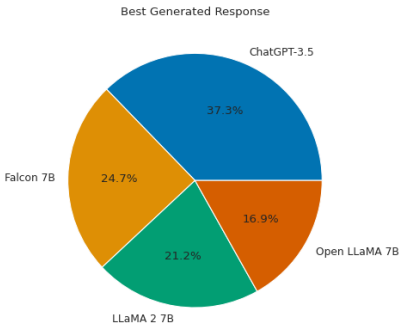
4.2 Human Evaluation

Given the insufficiency of previous metrics to estimate models' ability to generate specific responses to customer reviews, a human evaluation was developed to assess each language model's performance and facilitate comparison. This approach follows methodologies from prior studies on response generation to reviews [Cao and Fard 2021; Zhang et al. 2023]. For the evaluation, 50 negative customer reviews were randomly selected from the test dataset, with responses generated by each fine-tuned model and ChatGPT-3.5. These reviews were divided into 5 groups, each containing 10 reviews, resulting in 5 evaluation forms. To avoid bias, responses were not associated with model names. Respondents evaluated the quality of each response based on **grammatical fluency**, which measures how well the response is written and its ease of understanding; **informativeness**, which assesses the richness of information contained in the response; and **accuracy**, which analyzes how well the generated response accurately addresses the customer's complaint or issue. Each criterion is rated on a scale from 1 to 5, with higher scores indicating superior quality. After evaluating the responses, respondents determined which model generated the most appropriate response for each of the 10 reviews. Twenty-five volunteers completed the evaluation, with each form answered by 5 distinct volunteers, resulting in 250 evaluations. The compiled data calculated the average ratings for each criterion for each model. Additionally, the percentage distribution of choices for the best response was analyzed. Average ratings for Grammatical Fluency, Informativeness, and Accuracy are shown in Figure 2a, and the best response distribution is in Figure 2b.

To compare the models, a statistical test was conducted. A normality test (Kolmogorov-Smirnov) was performed due to the sample size being over 50 [P et al. 2019]. After confirming non-normal distributions, the Friedman test followed by the post-hoc Nemenyi test was conducted with a 5% significance level. To analyze grammatical fluency differences, we conducted the post-hoc Nemenyi test. Results show that ChatGPT-3.5 has significantly better fluency than LLaMA 2 ($p = 0.001$) and Open LLaMA 7B ($p = 0.004$). However, the difference between ChatGPT-3.5 and Falcon 7B was not statistically significant ($p = 0.059$), nor were there significant differences among the other models. This suggests that while ChatGPT-3.5 performs significantly better in fluency compared to some models, Falcon 7B is comparable in this criterion. In the Informativeness criterion, Falcon excelled among the developed models with an average score of 4.05, closely following ChatGPT-3.5's score of 4.17. This shows Falcon's effectiveness in incorporating customer review information into responses. The post-hoc Nemenyi test confirmed these results. ChatGPT-3.5 was significantly more informative than LLaMA 2 ($p = 0.028$) and Open LLaMA 7B ($p = 0.001$), but not significantly different from Falcon 7B ($p = 0.334$). These findings suggest ChatGPT-3.5 performs exceptionally well



(a) Average human evaluation metrics.



(b) Average human evaluation metrics.

Fig. 2: Human evaluation results.

in informativeness, but Falcon 7B is comparably effective in this criterion. Regarding Accuracy, Falcon and LLaMA 2 had close average ratings of 3.97 and 3.92, respectively, compared to ChatGPT-3.5's 4.19. This shows the fine-tuned models effectively address customer issues. The post-hoc Nemenyi test revealed ChatGPT-3.5 was significantly more accurate than Open LLaMA 7B ($p = 0.001$), but not significantly different from Falcon 7B ($p = 0.084$) or LLaMA 2 ($p = 0.054$). There were also no significant differences between Falcon 7B and the other models. These results indicate ChatGPT-3.5 is significantly better in accuracy than Open LLaMA 7B, but comparable to Falcon 7B and LLaMA 2. In choosing the best response for each comment, significant diversity was observed: ChatGPT-3.5 was the most chosen, followed by Falcon with 24.7%, LLaMA 2 with 21.2%, and finally, Open LLaMA with 16.9%. The results demonstrate that, although ChatGPT-3.5 performed better on the proposed criteria, the models we used in this study are capable of generating responses as effective as, or even better than, ChatGPT-3.5 in certain situations.

5. CONCLUSION AND FUTURE WORK

This article investigated the application of open-source language models in responding to customer reviews. The study included a literature review, selection of current open-source models with performance comparable to the state of the art, and their application to the proposed task. The research covered stages from extracting online customer reviews about restaurants to fine-tuning language models for responding to these reviews. The human evaluation revealed that the fine-tuned models, despite having significantly fewer parameters than ChatGPT-3.5, achieved comparable performance in terms of response richness and utility. ChatGPT-3.5 performed significantly better in grammatical fluency compared to LLaMA 2 and Open LLaMA 7B, but not Falcon 7B. In informativeness, ChatGPT-3.5 outperformed LLaMA 2 and Open LLaMA 7B, with Falcon 7B showing comparable performance. For accuracy, ChatGPT-3.5 was better than Open LLaMA 7B, but differences with Falcon 7B and LLaMA 2 were not significant. These findings suggest that fine-tuned models can generate responses nearly as rich and useful as those by ChatGPT-3.5, demonstrating the potential of pre-trained models enhanced through fine-tuning. Future work should involve fine-tuning pre-trained models with larger datasets, potentially exceeding 100,000 reviews, and optimizing hyperparameters for training efficiency [Cao and Fard 2021]. Additionally, maintaining coherence in longer dialogues and generating multiple responses for the same comment can improve natural interactions and model flexibility.

REFERENCES

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., et al. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023.

Alnuhait, D., Wu, Q., and Yu, Z. Facechat: An emotion-aware face-to-face dialogue framework, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Cao, Y. and Fard, F. H. Pre-trained neural language models for automatic mobile app user feedback answer generation. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*. IEEE, pp. 120–125, 2021.

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., and Brightwell, T. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*. Springer, pp. 3–17, 2023.

de Melo, T. Sentilexbr: An automatic methodology of building sentiment lexicons for the portuguese language. *Journal of Information and Data Management* 13 (3), 2022.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Farooq, U., Siddique, A. B., Jamour, F., Zhao, Z., and Hristidis, V. App-aware response synthesis for user reviews, 2020.

Gao, C., Zhou, W., Xia, X., Lo, D., Xie, Q., and Lyu, M. R. Automating app review response generation based on contextual knowledge. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31 (1): 1–36, 2021.

Lee, J., Park, D.-H., and Han, I. The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications* 7 (3): 341–352, 2008. Special Section: New Research from the 2006 International Conference on Electronic Commerce.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81, 2004.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology* 1 (2): 100017, Sept., 2023.

P, M., CM, P., U, S., A, G., C, S., and Keshri. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth*, 2019.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation, 2002.

Qing, P., Huang, H., Razzaq, A., Tang, Y., and Tu, M. Impacts of sellers' responses to online negative consumer reviews: Evidence from an agricultural product. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 66 (4): 587–597, 2018.

Qiu, H., He, H., Zhang, S., Li, A., and Lan, Z. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support, 2024.

Richardson, L. Beautiful soup documentation. *April*, 2007.

Sadiq, M. W., Akhtar, M. W., Huo, C., and Zulfiqar, S. Chatgpt-powered chatbot as a green evangelist: an innovative path toward sustainable consumerism in e-commerce. *The Service Industries Journal* 44 (3-4): 173–217, 2024.

Schwartz, S., Yaeli, A., and Shlomov, S. Enhancing trust in llm-based ai automation agents: New considerations and future challenges. *arXiv preprint arXiv:2308.05391*, 2023.

Shin, J., Tang, C., Mohati, T., Nayebi, M., Wang, S., and Hemmati, H. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks, 2023.

Sparks, B. A., So, K. K. F., and Bradley, G. L. Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management* vol. 53, pp. 74–85, 2016.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18 (6): 1–26, 2024.

Zhang, W., Gu, W., Gao, C., and Lyu, M. R. A transformer-based approach for improving app review response generation. *Software: Practice and Experience* 53 (2): 438–454, 2023.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2023.