Improving LLMs' Common Sense by Modeling Good Human Listeners

Joshua Tint **Edyssa Cervantes** **Rohitkumar Murali Arasanipalai** **Dipanshu Singh**

Abhik Chowdhury (Leader)
Arizona State University
achowdh5@asu.edu

Abstract

Transformer-based Large Language Models (LLMs) have achieved state-of-art results across a spectrum of tasks. However, they struggle to grasp nuanced emotional cues inherent in human communication, hindering their performance in emotional commonsense reasoning tasks. More emotionally aware LLMs have the potential open up future use cases in therapeutic domains, and would also help increase user trust and performance in existing chat domains. In this work we investigated if fine tuning models on chat interactions from professionally labeled "good" and "bad" listeners would improve or harm LLM's performance on emotional reasoning tasks in the SocialIQA dataset. We found this to be the case to varying degrees across the four models tested.

1 Introduction

Large Language Models (LLMs) have made remarkable strides in natural language understanding, achieving state-of-art results across a spectrum of tasks. However, a notable deficiency persists in their ability to grasp nuanced emotional cues inherent in human communication, hindering their performance in emotional commonsense reasoning tasks [3]. This limitation underscores the importance of addressing emotional awareness in LLMs, as they become increasingly ubiquitous in conversational interfaces and virtual assistants. An analysis of the attention distance, dispersion, and interdependency within LLM's by Jawale et. al. 2024 found that human conversations, relative to other forms of text on the internet such as code, or mathematical texts, require more nuanced handling of long-term contextual relationships and exhibit higher complexity. They emphasized the importance of training models to specifically handle human conversations. [2]. The Global Burden of Disease Study in 2010 discovered 10.4% of disability-adjusted life years stemmed from mental, neurological and substance use disorders. [8]. Training LLMs to become more emotionally aware could extend mental health access.

2 Literature Review

In recent years, researchers have explored various strategies to bolster LLMs' emotional reasoning capabilities. Chain-of-thought and prompting techniques have shown promise in nudging LLMs towards a better understanding of emotional contexts [7]. More emotionally aware LLMs potentially open up future use cases in therapeutic domains, and will help increase user trust and performance in existing chat domains [9]. Therefore, our work not only contribute to advancing the field of emotional reasoning in LLMs, but also helps in laying the ground work for real-world applications with benefits for end-users.

An overview of chatbots for mental health found that a 92.7% majority of investigated systems used only decision trees to generate responses and a 7.5% minority used any type of machine learning. [1] While rule based chatbots more reliably don't provide harmful information, they are less generalizable to the variety of possible client therapist interactions.

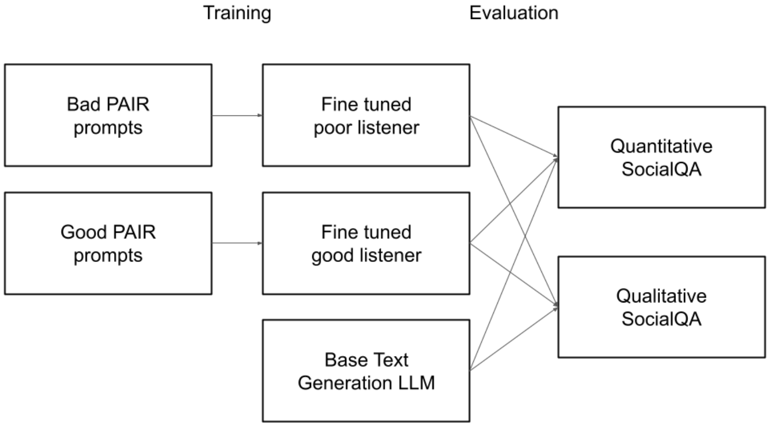

Figure 1: **A summary of our training and evaluation process per model**

3 Method

There were four LLMs examined in the experiment, including three chat models and one text completion model. The chat models were: Gemma7B, quantized at 4 bits, LLAMA7B, also quantized at 4 bits, and GPT3.5-Turbo. The text completion model was OpenAI's DaVinci-0002. These models were selected because of their popularity, to represent a diversity of model complexities, sizes, and origins.

In order to produce LLMs which modeled good human listeners, we used the PAIR dataset to fine tune. PAIR, which stands for Prompt-Aware margIn Ranking, is an expert-annotated chat dataset. Counselors annotated conversations as reflecting "good listening" and "bad listening," where good listening was defined as the ability to reflect on and empathize with information presented [5]. For each base model in our experiment, we fine-tuned a version on the good listening conversation transcripts and one on the bad listening transcripts in prompt-response format. For Davinci, we fine tuned a version using a prompt-completion format with "RESPONSE" tags separating each response in the conversation. For Gemma and LLAMA, the fine tuning was done across 2 epochs with a learning rate of `2e-04` and a weight decay of 0.001. For the OpenAI models, fine tuning was done across 3 epochs with a learning rate of `2e-04`, in line with default recommendations.

In order to test each LLM, we used the SocialIQA dataset, which has a broad variety of questions asking about commonsense reasoning as it pertains to emotions [6]. This includes questions asking to predict how a character's emotions would be influenced by phenomena, how a character's emotions would influence phenomena, and how a character's emotions would influence other emotions. We took a random sample of 500 questions from SocialIQA-test to evaluate our models with. SocialIQA is a closed ended dataset, with four options given per question. For each question, we prompted the LLM directly with the text of the question and the set of possible responses, each assigned a letter, then recorded whether the response indicated the letter of the correct answer. This allowed us to obtain closed-ended accuracies for each model's fine tuned and unmodified versions.

Grade	Description
0	There is no response or the response is not intelligible.
1	The response does not reference or mention any emotions.
2	The answer references a character's emotions but does not identify a causal relationship.
3	The answer identifies a line of reasoning involving emotions that is not reasonable.
4	The answer identifies a line of reasoning involving emotions which is reasonable.
5	The answer identifies multiple reasonable lines of reasoning involving characters' emotions.

Table 1: **Our coding rubric for open-ended results**

We also performed a qualitative analysis to better understand the relative strengths and weaknesses of each model. We took a random subset of 30 questions from SocialIQA, and each model was prompted with these questions. To get open-ended explanations rather than closed-ended responses, we modified each question to not give the list of possible responses and added a system prompt asking for a response with an explanation. We developed a 5-point rubric to evaluate each response based on what aspects of successful commonsense reasoning were present. Each member of our team graded a batch of responses such that each prompt was evaluated twice by two different graders. In order to validate our coding rubric, we used inter-coder agreement, setting a maximum disagreement threshold of 0.5 points. Our average disagreement was 0.42 points, thus we decided that our initial rubric did not need revisions for clarity.

4 Results

The closed-ended accuracies are shown in Table 2, alongside the average scores for the open-ended evaluations. For reference, the highest recorded closed-ended accuracy on the SocialIQA dataset is 83.2%, demonstrated by UNICORN-11B [4].

For our qualitative analysis, we recorded the average point score of each model and scaled them out of 100 as percentages to more easily compare them alongside closed ended accuracies.

Model Family	Fine Tune	Closed-Ended Accuracy	Open-Ended Score Percentage	Average Open-Ended Score
GPT3.5	good-listener	75%	62%	3.1
	bad-listener	52%	47%	2.3
	unmodified	77%	70%	3.5
Davinci	good-listener	35%	54%	2.7
	bad-listener	18%	43%	2.1
	unmodified	22%	2%	0.1
Gemma	good-listener	77%	82%	4.1
	bad-listener	69%	59%	2.9
	unmodified	76%	80%	4.0
LLAMA	good-listener	53%	84%	4.2
	bad-listener	34%	80%	4.0
	unmodified	44%	81%	4.1

Table 2: **The qualitative score of each model on the subset of the SocialIQA dataset**

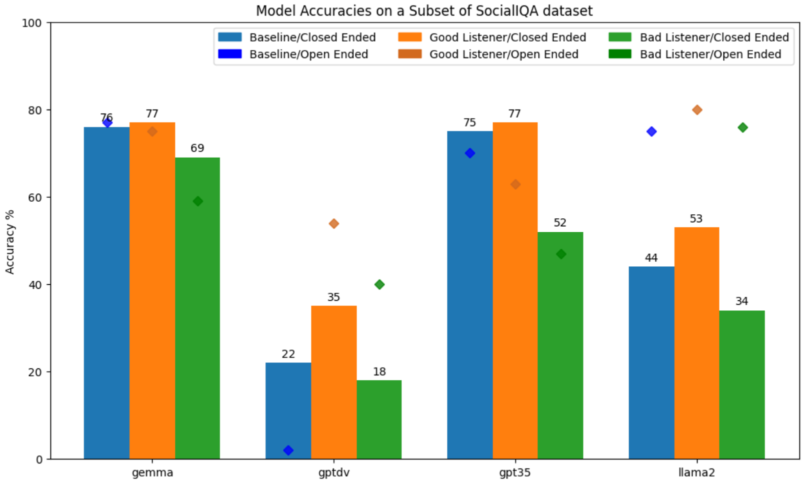

Figure 2: **The accuracies of each model in open- and closed-ended experiments**

5 Discussion

The most notable result is that every single "good listener" LLM outperformed its untrained counterpart on closed-ended accuracy, and every single "bad listener" LLM underperformed its untrained counterpart on closed-ended accuracy. Each gap was statistically significant at the p=0.05 level. These results clearly validate the use of fine tuning as a measure to improve emotional commonsense reasoning. However, the models with the lowest unmodified accuracies showed the biggest jumps after fine tuning; this show that fine tuning may be most effective on less-effective models.

Our qualitative results largely confirm and contextualize the results on closed-ended accuracy. Some results were unexpected, such as slight performance decreases after good-listener fine-tuning in Gemma and GPT3.5. However, these sample sizes were much smaller and not necessarily as representative of accuracies as the closed ended figures. Rather, they give us a better understanding of where each model falters. The best models in the open-ended experiments, Gemma and LLama, both consistently fail to expand their explanations to include multiple chains of reasoning, while the more commonly-used GPT3.5 more commonly struggles with identifying cause-effect emotional relationships that are reasonable to humans.

6 Future work

In this work due to computational limitations we tested relatively small versions of each model, quantized to 4 bits. Further research could explore the effect of fine tuning models that are larger in size on the PAIR data set, or other datasets annotated by listening quality. This would help confirm or reject the relationship between initial model performance and the marginal performance increase that fine-tuning brings. Because prompting methods have also been used to improve the emotional commonsense reasoning of models in the past, an experiment that explores the combination prompting methods and fine-tuning on emotional commonsense reasoning could be warranted. Additionally, because the PAIR dataset is relatively small, our results justify the creation of larger chat datasets annotated by listening quality.

References

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.*, 132(103978):103978, December 2019.

[2] Toshish Jawale, Chaitanya Animesh, Sekhar Vallath, Kartik Talamadupula, and Larry Heck. Are human conversations special? a large language model perspective, 2024.

[3] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[4] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark, 2021.

[5] Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[6] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms, 2023.

[8] Harvey A Whiteford, Alize J Ferrari, Louisa Degenhardt, Valery Feigin, and Theo Vos. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS One*, 10(2):e0116820, February 2015.

[9] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations, 2024.