Analysis of LLM's Ability to Produce Contextually Accurate Emotional Responses to User Selected Prompts.

NEILAN THUNBLOM, Colorado State University, United States
SEAN MURPHY, Colorado State University, United States

Large Language Models (LLMs) like OpenAI's GPT-3 and Google's Gemini have transformed the landscape of natural language processing, enabling the generation of responses that closely mimic human conversation. This study examines the efficacy of these models in delivering contextually accurate and emotionally attuned responses. By engaging two leading LLMs, this research evaluates their performance in responding to user-selected prompts that require nuanced emotional sensitivity. Participants rated responses based on emotional and tonal accuracy, providing insights into each model's ability to adapt to the emotional context of interactions. The results underscore significant advances in LLM capabilities but also highlight ongoing challenges in achieving consistency across diverse interaction scenarios. This study not only contributes to the understanding of LLMs' practical applications in complex communication tasks but also informs future enhancements needed to improve their empathetic responsiveness in user interactions.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Neilan Thunblom and Sean Murphy. 2024. Analysis of LLM's Ability to Produce Contextually Accurate Emotional Responses to User Selected Prompts. . *J. ACM* 37, 4, Article 111 (August 2024), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large Language Models (LLMs) have emerged as a powerful tool for media generation using natural language as an input. There is a rapidly growing number of applications that aim to utilize this new cutting-edge application of machine learning. Models such as OpenAI's GPT-3 and Google's Gemini are trained on large amounts of data pooled from across the internet, allowing them to generate human-like responses to questions asked by a user using natural language. Despite their wide application, the degree to which LLMs deliver context-sensitive and pertinent answers to user inquiries is still being explored [14].

The purpose of this study is to compare two LLMs and their ability to generate emotionally and tonally accurate responses to questions selected by the participant. These questions are not designed to elicit a specific response but are written to reflect a situation that a human would change their tone based on their personality to respond to. We compare the ratings of each response from OpenAI's GPT-3.5 model with the ratings of the same response from Google's Gemini Professional model. In doing this, we gained insight into the differences in how each model uses tone and emotion to emulate human traits in its responses. Recent research has demonstrated that even common NLP techniques

Authors' addresses: Neilan Thunblom, neilan@colostate.edu, Colorado State University, 1100 Center Ave Mall, Fort Collins, Colorado, United States; Sean Murphy, seanmurp@colostate.edu, Colorado State University, 1100 Center Ave Mall, Fort Collins, Colorado, United States.

Manuscript submitted to ACM

can be effectively applied to complex tasks such as codifying clinical notes, suggesting a broader applicability of NLP technologies in various domains [18]. Furthermore, a study by Jeong, Kim, and Kang (2023) has shown that multimodal integration in LLMs enhances emotional sensitivity and contextual understanding, which are crucial for tasks involving nuanced human interaction [8]. The findings from this study will inform the continued refinement of LLMs, specifically concerning the integration of personality and social awareness necessary for these models to provide responses that are on par with human emotional intelligence [4].

2 RELATED WORKS

Most recent studies on large language models are interested in their logical reasoning ability and how it relates to humans. The LLM's ability to detect tone and reason about it in order to provide a humanistic response accurately relates to its ability to reason. These studies have been conducted and will be referenced throughout our research on the abilities of these models to reason in a human way.

Significant research efforts have also been directed towards enhancing the capabilities of LLMs to handle tasks that require a nuanced understanding of text and context. For instance, Xiong et al. (2022) have explored the use of Multidimensional Latent Semantic Networks in recognizing text-based humor, an advanced application of discourse understanding. This study underscores the potential of LLMs to grasp subtle human expressions such as humor by analyzing semantic inconsistencies, phonetic features, and ambiguities, which are crucial for accurate humor recognition [19]. This aligns with the broader research trajectory aimed at making LLMs more responsive to the emotional and tonal subtleties of human language, enhancing their utility in user-centric applications.

Yang et al. advanced this field by proposing a multi-task learning system that not only optimizes for domain-specific responses but also enhances sentiment and aspect-aware summarization of reviews across different domains using deep reinforcement learning. This approach is particularly relevant for enhancing the emotional and tonal accuracy of responses generated by LLMs, as it focuses on adapting the model's output to align more closely with human expectations based on context, sentiment, and domain-specific content [20].

Overall the research and development of large language models has had a significant impact on the understanding of natural language and text in speech generation using natural language as an input. The study by Bang et al. introduced a new framework for evaluating interactive LLM's such as ChatGPT. This study was conducted using an evaluation of 23 datasets covering 8 different application tasks. The evaluation concluded that ChatGPT had a significant ability to multitask and could use multiple languages along with multi-model capabilities, as OpenAI advertises. This study also reveals the limitations of ChatGPT in generating non-Latin script languages along with its poor reasoning, which was stated to be around 63% across several categories of reasoning. Additionally, they found that ChatGPT suffers from hallucinations, as in many other models. However, these hallucinations can be very convincing and are due to the lack of understanding in memory due to limitations in the training data set.[3]

Another more recent study conducted on GPT 4 and its logical reasoning abilities involved testing the model using multiple logical reasoning datasets, including popular benchmarks like logiQA and Reclor. This study discovered that GPT 4 improved on chat GPT's performance, which, as stated previously, was around 63%. However, this study discovered that the model still struggles with out-of-distribution natural language inference datasets. This highlights the ongoing challenge of LLM's ability to reason logically, as discovered with ChatGPT and now with GPT 4.[11]

Recent research by Pieraccini et al. highlights advances in multimodal systems that integrate conversational interfaces and context-aware capabilities into vehicle systems, offering new insights into how LLMs could be optimized for interactive applications beyond typical text-based tasks [13]. This study, while focused on a specialized application,

contributes to our understanding of the broader implications of LLMs in handling real-world, dynamic interaction scenarios, further complicating the landscape of emotional and logical reasoning in AI systems.

Furthermore, Areshey and Mathkour demonstrated the potential of transfer learning with the BERT model to significantly enhance sentiment analysis tasks over traditional methods, emphasizing the advanced capabilities of LLMs in deriving contextual and emotional nuances from text, which is pivotal for applications requiring high-level understanding of user-generated content [2]. A similar study was published in 2023 which found that GPT3.5 is able to exhibit elements of empathy and in some cases was able to out preform humans in objective emathy metrics[17]

Another study conducted by Ratican and Hutson has highlighted the emotional intelligence of GPT 3.5 and GPT 4. This study is more in line with our research as it combines the emotional aspect of reasoning rather than primarily focusing on the logical aspect. The research emphasized the importance of developing models that have a comparable level of emotional intelligence to that of a human and that can provide more personalized and adaptive experiences[16]. Some of the things they studied are the models ability to offer empathetic support and promote the user's safety and independence. These are both aspects of our research study; however, we focus more on the tone and emotion given the prompt, while this study references a more broad statement involving the overall emotional dimension of multimodal LLMs. Furthermore, their research emphasizes the need to focus on emotional intelligence in order to ensure ethical and moral responses from the large language models that are being produced now in order for them to be able to maximize their benefits to humanity. [14][6]A recent study also discusses how these capabilities can be evaluated using newly developed methodologies, providing a comprehensive approach to assessing the nuanced capabilities of LLMs, including their potential biases and limitations in generating emotionally nuanced responses, especially in complex interactive settings [7].

3 METHODOLOGY

3.1 Experiment Design

In our study, we used a between-subjects design to evaluate the level of human-like emotion that was used in responses from the current leaders in large language models, Google Gemini and OpenAI GPT 3.5. We wanted to discover if there was a difference in the ability of these transformer-based LLMs to produce tonally accurate responses that reflect traditional human responses. We wanted to see if there was a difference in each model's ability to use the context of a question to produce responses that reflect human-like emotion and abide by social norms for conversation similar to [15]. For example, if someone told you they got hit by a car, the initial response should not be to offer solutions or blame but instead to sympathize with the person and ask how they are doing. While it is not grammatically or logically incorrect for you to say something along the lines of "Why were you in the street? That's just asking to get hit", a more human-like, emotional-based response would be, "I'm sorry that happened to you. Is there anything I can do to help?" Our goal was to determine if these LLMs had the ability to respond in a tonally correct manner to a wide variety of prompts and, if so, if either of the leading models did a better job. Ultimately we believe that LLMs can be used in mental health and customer service settings where emotional intelligence that meets or exceeds the level of a human is necessary for positive outcomes[9].

Both the OpenAI GPT and Google Gemini Pro models were tasked with responding to a set of pre-selected questions designed to incite different levels of emotion in their responses, similar to how a human would respond. The participants in our study were then asked to rate the LLM's response to the selected prompt in terms of tonal and contextual accuracy. We assigned participants to one of two groups representing the model that they would be using.

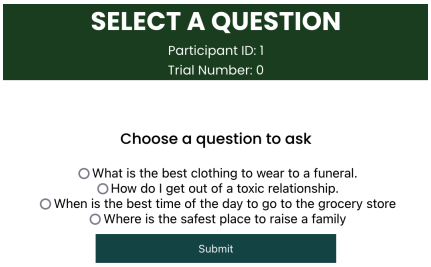

Fig. 1. The Prompt Selection Page

3.2 Application Design

Our experiment uses a web application to gather data from participants about the level of emotion in response to a question asked to an LLM. We used this platform as an interface for the user to interact with the LLM via its respective API. The application allowed users to select questions, send them to the assigned LLM and then receive the response to evaluate. We used a React front-end to design the user interface and workflow and a Node.js back-end to handle the data logging and the calls to the respective LLM API. This application was designed to be user-friendly and with accessibility in mind, using accessible fonts and colors as determined by Colorado State University. It was also designed to abstract away the technological aspects of interacting with LLMs in order to reduce bias in responses.

The application consists primarily of 3 pages. The first being the consent page where the experiment and its implications are explained to the participant. Here, the participant is assigned an ID and a group. If they agree to the consent statement, the question page is displayed, where the participant chooses from one of four randomly selected prompts from our prompt bank. Upon clicking the submit button as shown in figure 1 they are shown a loading screen while the server waits for a response from the selected LLM's API. Once the response returns, the review page is shown asking the participant to select the level of emotional accuracy on a scale from 0 to 5 as shown in 2.

3.3 Data Collection

After filling out the content form, the participant selected a prompt and rated the response over 10 trials using a different question for each trial, pulled from the prompt bank. The application back-end would log the id of the participant, their group number and either the prompt or the response along with the users rating, depending on the stage in the workflow the user was in. This process allowed us to ensure that each of the participants' data was recorded in a structured way, allowing for easy analysis later in the process.

This data is then fed into a Python script that maps the prompt from each trial to a key and appends the rating to the end. This format allows us to run the statistical methods on the data in an easy-to-manage way.

3.4 Analysis

The ratings we recorded from the study were analyzed using analysis of variance; this statistical method allowed us to determine the level of statistical significance between the ratings of responses from each of the LLMS in the context of tonal and contextual accuracy. This was a significant analysis for our experiment as it enabled us to assess if one model

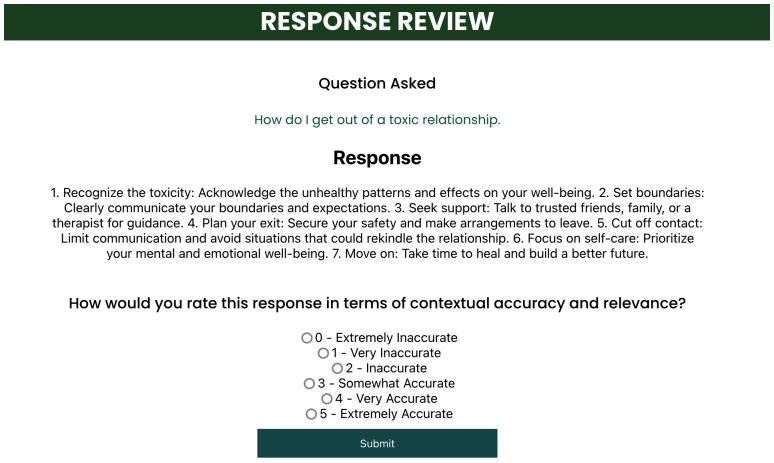

Fig. 2. The Response Rating Page

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Google Gemini	55	242	4.4	0.9111
OpenAI GPT	55	237	4.309	1.2175

Fig. 3. Summary of Responses

did better than the other in producing tonally accurate responses that resemble that of the average human. The Anova, along with the summary statistics, were computed using Microsoft Excel and analyzed for discussion below.

4 RESULTS

Overall, our findings were that each LLM did an excellent job of producing responses to the prompts in a tonally and emotionally accurate manner. We discovered that both of the models had similar behavior and were both able to produce responses that were deemed accurate by participants, according to our metrics collected3. As shown in figure 3, the Gemini model had an average response rating of 4.4, which is between "Very Accurate" and "Extremely accurate" on the scale used by participants. Similarly, the GPT 3.5 model had an average response rating of 4.3, putting it in the same "Very Accurate" and "Extremely Accurate" scale but leaning towards the lower end. Using just the average of the ratings suggests that both models do a very good job of responding to prompts with a tone that is expected based on the context of the question.

Looking into the ANOVA shown in figure 4 allows us to see additional information about the differences between the groups and how they relate to our research question. The F statistic we achieved was 0.213, which is significantly lower than the Critical F value of 3.929. This indicated that there is not a statistically significant difference between the models in the tone they utilize in their responses. This lack of statistical significance is reinforced by the large p-value of

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.2273	1	0.2273	0.21354	0.6449	3.92901
Within Groups	114.9455	108	1.0643			
Total	115.1727	109				

Fig. 4. ANOVA

0.64, indicating that there is not that much of a difference in each model's ability to produce a tonally accurate response to questions it is asked.

5 DISCUSSIONS

The study's results indicate that both OpenAI's GPT-3.5 and Google's Gemini are adept at producing responses that align closely with human expectations in terms of emotional and contextual appropriateness. While not statistically significant, the slight edge in average response ratings for Gemini suggests a minor variation in performance that does not substantively differentiate the two models. This observation implies that both LLMs are converging towards a high standard of emotional intelligence, challenging the existing benchmarks set within the field.

The proficiency of these models in generating emotionally nuanced responses underscores their potential utility in applications requiring sensitive interaction, such as in customer support or mental health advisement. The successful deployment of the Psy-LLM framework in providing effective psychological support demonstrates the practical application of such technologies, offering a blueprint for the future utilization of LLMs in mental health services [10]. Research by Amirova et al. into the algorithmic fidelity of LLMs illustrates the critical role these models play in understanding and mimicking human emotional responses, particularly in complex sociocultural contexts [1]. Additionally, Masoumi et al.'s studies further support the importance of algorithmic fidelity, emphasizing the potential for these models to achieve nuanced understanding in varied applications [12]. Moreover, the study by Bouazizi et al. (2024) extends this discussion by highlighting the intricate challenges involved in deploying LLMs for nuanced emotional recognition, such as detecting subtleties in mental states from speech in diverse cultural contexts, where traditional models may overlook critical emotional cues [5]. However, the absence of a significant difference also highlights a possible ceiling effect in the current generation of LLM technologies, where incremental improvements may require new innovations or methodologies in model training and architecture, such as a move to diffusion-based models from these transformer-based models.

Future research could address the limitations noted in the current study by broadening the emotional range of the prompts used and incorporating a more diverse demographic profile of participants to test variations in emotional perception and response accuracy. Moreover, longitudinal studies could provide insights into the consistency and reliability of these models over extended interactions, which are typical in real-world applications. A comparative analysis with human responses would also help in benchmarking these models against genuine human emotional intelligence, offering a more nuanced understanding of where these models stand and where they might still be lacking.

6 CONCLUSION

This research confirms the capability of leading LLMs like GPT-3.5 and Gemini to generate contextually suitable and emotionally resonant responses. The findings suggest that both models perform comparably well, with no significant statistical difference in the quality of their responses. This result points to the advanced state of current LLMs but also

to the challenges ahead in surpassing the existing levels of performance. Innovations in training data diversity, model architecture, and emotional intelligence metrics will be crucial for the next leap forward in LLM development. Such advancements will enhance the practical deployment of these models in fields where understanding and responding to human emotions are critical, ultimately leading to more natural and effective human-computer interactions.

REFERENCES

[1] Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed, Martin R. Cowie, and Joel Z. Leibo. 2024. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. *PLoS ONE* 19 (Mar 2024), p.1–33. Issue 3.

[2] Ali Areshey and Hassan Mathkour. 2023. Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model. *Sensors* 23 (Jun 2023), p5232. Issue 11. https://doi.org/10.3390/s23115232

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and et al. 2023. A multitask, multilingual, multimodal evaluation of Chatgpt on reasoning, hallucination, and Interactivity. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023). https://doi.org/10.18653/v1/2023.ijcnlp-main.45

[4] Lisa Feldman Barrett. 2017. The theory of Constructed Emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* 12, 11 (May 2017), 1833–1833. https://doi.org/10.1093/scan/nsx060

[5] Mondher Bouazizi, Chuheng Zheng, Siyuan Yang, and Tomoaki Ohtsuki. 2024. Dementia Detection from Speech: What If Language Models Are Not the Answer? *Information* 15 (Jan 2024), p2–22. Issue 1. https://doi.org/10.3390/info15010002

[6] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[7] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2024. Who Determines What Is Relevant? Humans or AI? Why Not Both? *Commun. ACM* 67 (Apr 2024), p31–34. Issue 4. https://doi.org/10.1145/3624730

[8] Eunseo Jeong, Gyunyeop Kim, and Sangwoo Kang. 2023. Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information. *Mathematics* 11 (Jul 2023), p2908. Issue 13. https://doi.org/10.3390/math11132908

[9] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. arXiv:2307.11991 [cs.CL]

[10] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2024. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics* 4, 1 (Mar 2024), p.8–33. https://www.biomedinformatics.com/issues/march2024/articles/supporting-the-demand-on-mental-health-services-with-ai-based-conversational-large-language-models-llms

[11] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of CHATGPT and GPT-4. https://arxiv.org/abs/2304.03439

[12] Safoora Masoumi, Hossein Amirkhani, Najmeh Sadeghian, and Saeid Shahraz. 2024. Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. *Systematic Reviews* 13 (Apr 2024), p1–9. Issue 1. https://doi.org/10.1186/s13643-024-02470-y

[13] Roberto Pieraccini, Krishna Dayanidhi, Jonathan Bloom, Jean-Gui Dahan, Michael Phillips, Bryan R. Goodman, and K. Venkatesh Prasad. 2004. Multimodal Conversational Systems for Automobiles. *Commun. ACM* 47 (Jan 2004), p47–49. Issue 1. https://doi.org/10.1145/962081.962104

[14] Jay Ratican and James Hutson. 2023. The Six Emotional Dimension (6de) model: A multidimensional approach to analyzing human emotions and unlocking the potential of Emotionally Intelligent Artificial Intelligence (AI) via Large Language Models (LLM). *Digital Commons@Lindenwood University* (May 2023). https://digitalcommons.lindenwood.edu/faculty-research-papers/481/

[15] Ciaran Regan, Nanami Iwahashi, Shogo Tanaka, and Mizuki Oka. 2024. Can Generative Agents Predict Emotion? arXiv:2402.04232 [cs.AI]

[16] Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. 2023. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* 15, 6 (2023). https://doi.org/10.3390/fi15060192

[17] Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large language models (llms) and empathy – a systematic review. https://www.medrxiv.org/content/10.1101/2023.08.07.23293769v1

[18] Nazgol Tavabi, Mallika Singh, James Pruneski, and Ata M. Kiapour. 2024. Systematic evaluation of common natural language processing techniques to codify clinical notes. *PLoS ONE* 19 (Mar 2024), p1–13. Issue 3. https://doi.org/10.1371/journal.pone.0298892

[19] Siqi Xiong, Bongbo Wang, Xiaoxi Huang, and Zhiqun Chen. 2022. Multidimensional Latent Semantic Networks for Text Humor Recognition. *Sensors* 22 (Aug 2022), p5509–N.PAG. Issue 15. https://doi.org/10.3390/s22155509

[20] Min Yang, Qiang Qu, Ying Shen, Kai Lei, and Jia Zhu. 2020. Cross-domain aspect/sentiment-aware abstractive review summarization by combining topic modeling and deep reinforcement learning. *Neural Computing Applications* 32 (Jun 2020), p6421–6433. Issue 11. https://doi.org/10.1007/s00521-018-3825-2

Received 30 March 2024; revised 28 March 2024; accepted 30 March 2024