An Examination of "Empathy" in Conversational Interfaces for LLMs

Alok Debnath*
ADAPT Centre,
School of Computer Science and Statistics,
Trinity College Dublin
debnatha@tcd.ie
alokdebnath.github.io

Allison Lahnala
Independent Researcher
Hamilton, Ontario, Canada
alahnala@gmail.com
alahnala.github.io

1 Extended Abstract

Large Language Models (LLMs) are ubiquitous today and several of the commercially viable models use a conversational interface (e.g., ChatGPT) for interacting with the user (Kalla et al., 2023). User-facing language models are now expected to be multifunctional—fluent, relevant, informative, and emotionally aware—blurring the lines between previously independent tasks with distinct datasets, architectures, and evaluations, as these multifunctional capabilities are increasingly treated like traditional tasks as models become more advanced (Fomicheva et al., 2021; Ni et al., 2023).

Empathetic response generation has become one such task for LLMs. Empathy is a desirable trait for its supposed functions in increasing comfort, user-agent bonding, improving user trust, and building understanding of emotional intents in conversations (Chen et al., 2023), among other hypothesized benefits. Rather than a mechanism that *could* underly or a variable that *may* affect the primary outcome, however, empathy is often thought to be *the* target outcome, driving attempts to measure it in language to evaluate generated empathy. Our work argues that the ambiguity of defining empathy makes it challenging to pin down, and suggests that understanding empathy should involve a context-specific analysis of the domain, role, and situation in which an agent interacts with the user.

The study of empathy broadly examines individuals' behaviours and cognition patterns during interactions in a given context (Clark et al., 2018). However, the variability in the contexts can make it a nebulous concept to operationalize (Cuff et al., 2016). This is emphasized in NLP in two ways. First, any interpretation of "behaviours" and "cognitive patterns" has to be constrained within the modality of the task at hand while also aligning with the annotation guidelines/schema, dataset, model architecture, evaluation framework and auxiliary tasks (Lahnala et al., 2022). Second, any adopted schema or definitions might only be contextually applicable and therefore would not be comparable when switching domains (e.g. from mental health to customer service) or generalizing (e.g. from mental health to the open domain) (Cuadra et al., 2024).

These contextual differences are emphasized in a conversational interface where, metaphysically, *how do we tell if a conversational agent is empathetic*? For LLMs, any model's inherent capability to be a fluent generator of a language can not be wholly evaluated, only each production can be rated on a constrained scale in the context of that production, analogous to the idea of communicative competence (Hymes, 1992). Studying the constraints and expressive functions of conversational interfaces can create a clearer framework for evaluating emotional intelligence, common-sense reasoning, and intent understanding by aligning observable properties with specific expectations.

Traditional considerations about information privacy and user safety have also become more pressing. Concerns about data privacy and security often revolve around the LLM capturing, processing, and storing sensitive user data as training material (Li et al., 2024). These threats are now compounded by the model's ability to pass the Turing test, be persuasively human, and perform social engineering with little oversight (Ai et al., 2024). Now, data security is a threat even for those not interacting with the LLM even by virtue of its presence on the internet. Such a threat is pervasive and its scale is unprecedented.

Lastly, there is the question of why. Why should we require agents to mimic human empathy when no universally accepted definition of empathetic behavior exists? Emotions manifest differently in language for myriad reasons, cultural, social, contextual, and personal, and there is no global average baseline degree of empathy that is exhibited or

posited through language. In fact, there are several reasons not to imbue agents with empathy, whether for the culturally ill-informed assumption that there exists a gold standard of empathetic interaction or the possibly ethically dubious (or downright nefarious) applications of an empathetic, persuasive, and unconscious computational conversant (e.g. scamming the technologically unaccustomed (Distler et al., 2023), reaffirming discriminatory ideologies, and purporting human-like companionship). Overall, we call for further research into *operationalizable* studies of empathy before rushing headlong into humanising LLMs.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, et al. 2024. Defending against social engineering attacks in the age of llms. *arXiv preprint arXiv:2406.12263*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Malissa Clark, Melissa Robertson, and Stephen Young. 2018. "i feel your pain": A critical review of organizational research on empathy. *Journal of Organizational Behavior*, 40.

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 *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Benjamin Cuff, Sarah Brown, Laura Taylor, and Douglas Howat. 2016. Empathy: A review of the concept. *Emotion Review*, 8:144–153.

Verena Distler, Yasmeen Abdrabou, Felix Dietz, and Florian Alt. 2023. Triggering empathy out of malicious intent: the role of empathy in social engineering attacks. In *Proceedings of the 2nd Empathy-Centric Design Workshop*, pages 1–6.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178.

Dell Hymes. 1992. The concept of communicative competence revisited. *Thirty years of linguistic evolution*, 1(2):31–57.

Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. 2024. Llm-pbe: Assessing data privacy in large language models. *Proceedings of the VLDB Endowment*, 17(11):3201–3214.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.