Modelling Human-like Emotions with Generative Agents

Ciaran Regan[1], Nanami Iwahashi[1], Shogo Tanaka[2] and Mizuki Oka[1]

[1]Grad. School of Science and Technology, University of Tsukuba, Japan
[2]Grad. School of Letters, Tokai University, Japan
mizuki@cs.tsukuba.ac.jp

Abstract

Recent studies suggest that Large Language Models (LLMs) can achieve similar performance to humans on some Theory of Mind (ToM) tasks. However, LLMs are unable to fully mimic human-like empathy. In this work, we investigate the ability of LLMs to predict emotion in various situations by introducing a novel agent architecture in which new experiences are compared to past memories via a norm. Through this comparison, agents gain the ability to comprehend new experiences in context, which according to the appraisal theory of emotion, is vital in emotion creation. By describing a variety of experiences in natural language, we test the emotional responses of agents across a wide range of scenarios. The mixed results suggest that although the norm improves the ability to mimic empathy, LLMs still struggle to capture negative emotions. These findings underscore the need for further research into techniques for aligning the emotional intelligence of generative agents. Ultimately, this work takes a step towards more empathetic and socially aware AI systems, which could have significant implications for human-AI interaction and collaboration.

Introduction

Large Language Models (LLMs) have exhibited a number of emergent abilities (Wei et al., 2022). Although there is significant debate, some argue that Theory of Mind (ToM) may have emerged in LLMs as a byproduct of their improved language skills (Strachan et al., 2024; Kosinski, 2023). ToM, which is defined as the ability to impute unobservable mental states of others, enables humans to track the emotions, intentions, beliefs and desires of third parties and is thought to play a key role in social interactions, communication and empathy. Given the importance ToM plays in human interactions, there have been significant efforts to equip AI with ToM-like abilities to achieve a more safe and human-like AI (Yang et al., 2018).

Although able to react appropriately in certain situations, LLMs fall short in alignment with the emotional behaviours of humans and cannot establish connections between similar situations (Huang et al., 2023). One of the possible explanations for this is that a LLM cannot respond to events in the same way as humans due to the lack of criteria to assess them that have been formed through related episodic memories. On the one hand, according to the appraisal theory of emotions (ATE), a cognitive approach to understanding emotions, our appraisals of the significance of the event triggers and determines a proper emotion in the given environment (Scherer, 1999; Moors et al., 2013). That is, how we assess events directly influences how we emotionally respond to them. On the other hand, neuropsychology suggests that episodic memories shape how we perceive new events (Baddeley, 1982). Based on the memories of past experiences, our brain generates a model of the world around us that informs our perception of upcoming events (Zeidman and Maguire, 2016). In this regard, the role of episodic memory seems crucial in generating both the criteria to assess an event and the model to perceive new events. Although LLMs can posit a guess of the emotions an experience would cause due to their vast amounts of training data, they lack episodic memory, which is required by ATE to accurately simulate human-like emotional responses.

Agent Architecture

Based on (Park et al., 2023; Regan et al., 2024), the following architecture is implemented, depicted in Fig 1. The agent receives experiences through natural language, which acts as perception. Past memories are then retrieved, weighted by saliency, relevancy and recency. These memories are summarised into what is referred to as the "norm", which is designed to capture insights such as the agent's habits and expectations.

Subsequently, the norm is compared to the new experience to create a "contextual understanding", which captures the differences between the current situation and the established norm. To assess the emotional response of the agent, the Positive and Negative Affect Schedule (PANAS) is administered, in which the agent rates their affect level in a variety of positive and negative emotions (Watson et al., 1988), allowing the agent's emotional response to be probed[1]. Finally, the new experience is stored as a memory, which can be utilised in the creation of future norms.

[1]Although referred to as an emotional response throughout this work, we note that this only a superficial mimicking of emotion.

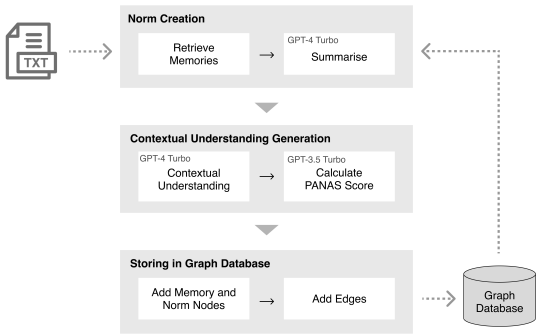

Figure 1: The proposed architecture.

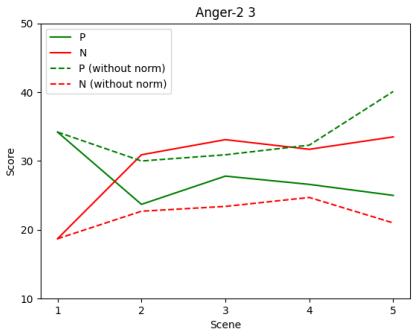

Figure 2: PANAS scores for EmotionBench "Anger-2 3" with/without norms, as a 5-part story.

Experiment

To analyse how the emotional response evolves, a dataset of 5-scene, emotionally neutral stories was created by expanding the scenarios from EmotionBench (Huang et al., 2023) using OpenAI's GPT-4. These scenes play the role of episodic memories for the agent. For each scene in the story, the architecture is run 10 times and the agent's average emotional response is measured. The architecture is also tested without the norm component to analyse how context influences the emotional state.

Results

Plotting the evolution of PANAS scores reveals that there are situations in which the architecture is effective and ineffective for emotional alignment with humans. Consider the following experiences where the architecture qualitatively improved alignment, with the PANAS results shown in Fig 2.

1. I am spending time in the living room with my two brothers when a disagreement begins.
2. As we exchange words, the situation develops into a physical one, and I receive a hit in the abdomen.
3. Following the hit, I instinctively react with a physical response directed at both of my brothers.
4. Upon my reaction, my brothers increase the intensity of their physical actions in the dispute.
5. The physical exchange between us persists, and there are no parents present to intervene.

Initially, the scores with and without norm are identical. Following this, the second experience triggers a strong negative reaction with the proposed architecture. This can be attributed to the agent understanding that there is an escalation of a family conflict, as described by the "contextual understanding" at that moment: "The situation is a red flag that the family might need to address the way disagreements are handled to promote a safer, more supportive family environment."

In cases of ineffective alignment, ambiguous situations were found to be interpretable in various ways, despite the provided context. In particular, GPT-3.5-Turbo tends to report high positive affect in unclear situations.

Table 1 shows the average change in PANAS scores across all stories, highlighting the difference between the agent's current PANAS score and its baseline score (when no experience is provided). Given EmotionBench uses only negative emotions, the minimum positive and maximum negative scores from each of the five experiences are averaged.

	With Norm		**Without Norm**	
	P (min)	**N (max)**	**P (min)**	**N (max)**
Δ Affect	$\downarrow$ (-18.0)	$\uparrow$ (+1.6)	$\downarrow$ (-18.7)	$-$ (+0.3)

Table 1: The average change in Positive (P) and Negative (N) affect for all stories in the dataset.

Conclusion and Limitations

The results show that the norm allows for a greater increase in negative affect, suggesting the additional context improves emotional alignment. However, the decrease in positive affect is significantly greater than the increase in negative affect, in agreement with Huang et al. (2023), which found that GPT-3.5-Turbo fails to react appropriately in negative situations.

While these results demonstrate the importance of episodic memories, this work faces limitations, such as comparing the responses to that of humans. Furthermore, it would be beneficial to study a wider range of scenarios, such as the response to non-sensical situations. Nevertheless, this work takes a step towards more empathetic AI, which could have significant implications for human-AI interaction.

Acknowledgment

This work was supported by MIXI, Inc.

References

Baddeley, A. D. (1982). Implications of neuropsychological evidence for theories of normal memory. *Philos Trans R Soc Lond B Biol Sci*, 298(1089):59–72.

Huang, J.-t., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z., and Lyu, M. R. (2023). Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Kosinski, M. (2023). Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pages arXiv–2302.

Moors, A., Ellsworth, P., Scherer, K., and Frijda, N. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5:119–124.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Regan, C., Iwahashi, N., Tanaka, S., and Oka, M. (2024). Can generative agents predict emotion? *arXiv preprint arXiv:2402.04232*.

Scherer, K. R. (1999). Appraisal theory. In Dalgleish, T. and Power, M., editors, *Handbook of cognition and emotion*, pages 637–661. Wiley, Chichester, UK.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., et al. (2018). The grand challenges of science robotics. *Science robotics*, 3(14):eaar7650.

Zeidman, P. and Maguire, E. A. (2016). Anterior hippocampus: the anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience*, 17(3):173–182.