Demographic Prompting Fails to Bridge the Individual Variability Gap: GPT-4o Aligns with Average but Not Individual Emotional Ratings of Images

Chace Ashcraft (Chace.Ashcraft@jhuapl.edu)
11100 Johns Hopkins Road
Laurel, Maryland 20723 USA

Raphael Norman-Tenazas (Raphael.Norman-Tenazas@jhuapl.edu)
11100 Johns Hopkins Road
Laurel, Maryland 20723 USA

Rik Bose (Rik.Bose@jhuapl.edu)
11100 Johns Hopkins Road
Laurel, Maryland 20723 USA

Michael Wolmetz (Michael.Wolmetz@jhuapl.edu)
11100 Johns Hopkins Road
Laurel, Maryland 20723 USA

Mattson Ogg (Mattson.Ogg@jhuapl.edu)
11100 Johns Hopkins Road
Laurel, Maryland 20723 USA

Abstract

Large language models (LLMs) and vision language models (VLMs) have been shown to closely align with human behavior in aggregate, but tend to align less well with individuals, and poorly approximate the variability of cohorts of human agents. We explored aligning models to specific individuals based on their demographic data on an emotion rating task by eliciting ratings along two standard psychological emotion dimensions on the previously human-normed OASIS dataset. We created AI "proxy" participants for human participants in the original OASIS study by prompting GPT-4o with a human participant's demographic data, then instructed the AI participant to rate a set of images for emotional valence or arousal, reproducing the human paradigm. We found that group-averaged GPT-4o ratings correlated to group-averaged human responses, but observed different distributions of responses. Representations of specific individuals poorly aligned with human ratings, despite using specific demographic data. In general, GPT-4o appears to align fairly well with human emotional responses on average, but work is needed to capture human variability to enable VLMs to emulate the behavior of specific individuals.

Keywords: vision language model; large language model; emotion; alignment

Introduction

Every successive generation of large language models (LLMs) and vision language models (VLMs) demonstrates increasingly impressive human-like capabilities for communication (Jones & Bergen, 2024, 2025), reasoning (Minaee et al., 2025) and perception tasks (Tiganj, Dickson, Maini, & Nosofsky, 2025; Marjieh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2024; Ogg, Bose, Scharf, Ratto, & Wolmetz, 2025). However, their understanding of, or alignment with, human emotion is less well understood (He, Guo, Rao, & Lerman, 2024; Chang, 2024; Sabour et al., 2024). Emotions are complex phenomena that play an important role in everyday life (Pessoa, 2008), and assuming that artificial intelligence (AI) accurately represents emotional responses in humans could lead to poor performance in critical roles or friction with human users.

At the same time, numerous results have shown that LLMs and VLMs produce homogeneous responses that do not capture the variability found within cohorts of human participants (Abdurahman et al., 2024; Mei, Xie, Yuan, & Jackson, 2024; Ogg et al., 2025; He et al., 2024). Methods are beginning to be developed to address this problem (Castricato, Lile, Rafailov, Fränken, & Finn, 2024), and there has been some encouraging progress in aligning foundation models with individual human data (Zhao et al., 2025), but this remains a central problem for many applications of LLMs and VLMs, where they might be used to represent or simulate human behavior.

Methods

To investigate the similarity of AI emotional judgments with human ratings of different emotions at the group and individual level, we used a dataset of images for which humans reported their emotional reactions along the standard psychological dimensions of valence and arousal (Russell, 1980; Kensinger, 2004). Specifically, we used the images and demographic data provided in the Open Affected Standardized Image Set (OASIS) dataset (Kurdi, Lozano, & Banaji, 2017). We constructed AI *proxies*, or representative stand-ins, of each human participant in the OASIS human study by providing a description of their demographic data to GPT-4o (OpenAI et al., 2024). For a given participant, a prompt is constructed, instructing GPT-4o to represent a person of the demographic background of the participant, after which the GPT-4o proxy is asked to rate the same set of images the participant rated in the OASIS study, allowing us to compare emotional LLM ratings of images directly with the participant it was intended to represent by proxy.

The OASIS dataset consists of 900 color images spanning various themes, in particular: humans, animals, objects, and scenes. The human study in the introductory OASIS work included 822 human subjects. Subjects were divided into groups, one rating images for emotional valence and the other emotional arousal (Kensinger, 2004), allowing participants to focus on a single psychological measure. Participants were taught the intended meaning of their assigned measure and then instructed to rate 225 images sampled from the 900 on a Likert scale of 1-7.

We selected 78 participants from the OASIS study. For each participant, we construct a proxy initialization prompt using their demographic information. The prompt is provided to GPT-4o, telling it that it is a person of that background, even giving it a fictional name. For example:

> "You are Scot Hoover, a 29 year-old man with moderately liberal political leanings. You consider yourself to be White. You have a college degree education and your current household income is below $25,000 per year."

GPT-4o is also given a 150x150 pixel version of the image and instructions to rate, it following the protocol of the OASIS study as closely as possible. Some images, such as sexually explicit or violent images, were skipped because they violated GPT-4o's content policies.

Results and Conclusion

Figure 1 shows that while the proxy participants' scores better cover the range of response options, they tended to be concentrated around 4 for valence, and occupied a bimodal distribution (between 3 and 4, and between 5 and 6) for arousal, whereas human ratings were less concentrated at certain values. Overall, we observe a correlation of $r = 0.8$ ($p < 0.001$ for all correlations) between the human and GPT-4o proxy ratings of images, suggesting that, when averaged across participants, the proxy responses strongly correlate with the human responses at the group level.

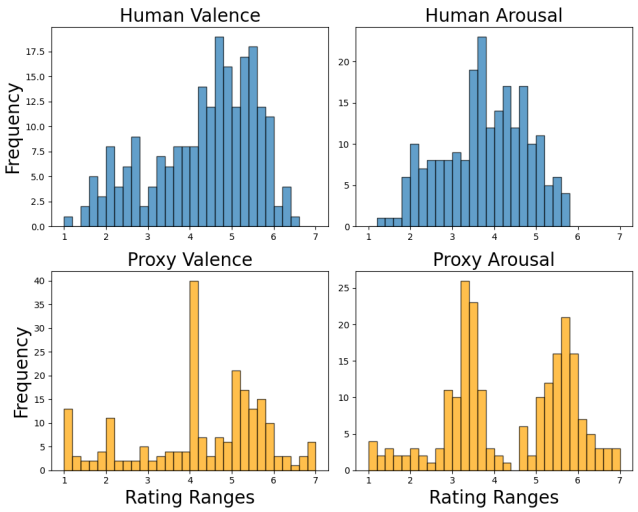

Figure 1: Valence (left) and arousal (right) ratings for human (top) and GPT-4o proxy (bottom) participants averaged for each image.

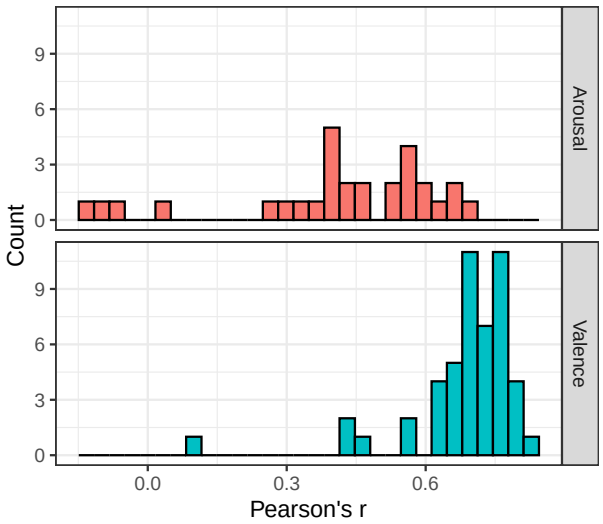

Figure 2: Distribution of alignment observed between individual human raters and their corresponding AI proxies for arousal (top) and valence (bottom).

However, at the individual level, alignment between specific human raters and corresponding GPT-4o proxy raters was lower, suggesting a weaker encapsulation of the responses of specific individuals (median individual-level human to AI Proxy correlation is $r = 0.44$ and $r = 0.71$ for arousal and valence, respectively; Wilcoxon rank sum test $W = 119,\ p < 0.001$). Figure 2 shows that this is particularly true for arousal. Figure 3 shows the distribution of human-to-human correlations and proxy-to-proxy correlations, where it is clear that the proxy participants' responses are highly correlated with each other, while the human participants' responses are more heterogeneous. We confirmed that initializing AI proxy participants using demographics slightly increased the heterogeneity of rater responses, compared to ratings without individualized prompts based on demographics ($W > 5770,\ p < 0.001$ for both arousal and valence ratings), but that overall, heterogeneity of AI proxy responses was quite high relative to human responses ($W > 123201,\ p < 0.001$).

We conclude that in cases where emotion or individuality are important factors, current VLMs, such as GPT-4o, are likely to be poor human proxies given the prompting strategies studied here. Those attempting to leverage AI as a proxy for human experience – e.g. (eun Yoon, He, Echterhoff, & McAuley, 2024; Ziv, Lan, Chemla, & Katzir, 2025) – may be able to approximate an average emotional sentiment of a diverse population through modern LLMs and VLMs, but they will likely fail to represent specific individuals or produce the variability seen in human data (also seen in (He et al., 2024). Future efforts might explore deeper prompting and proxy initialization strategies, or could develop methods for fine-tuning a model's behavior to represent specific individuals.

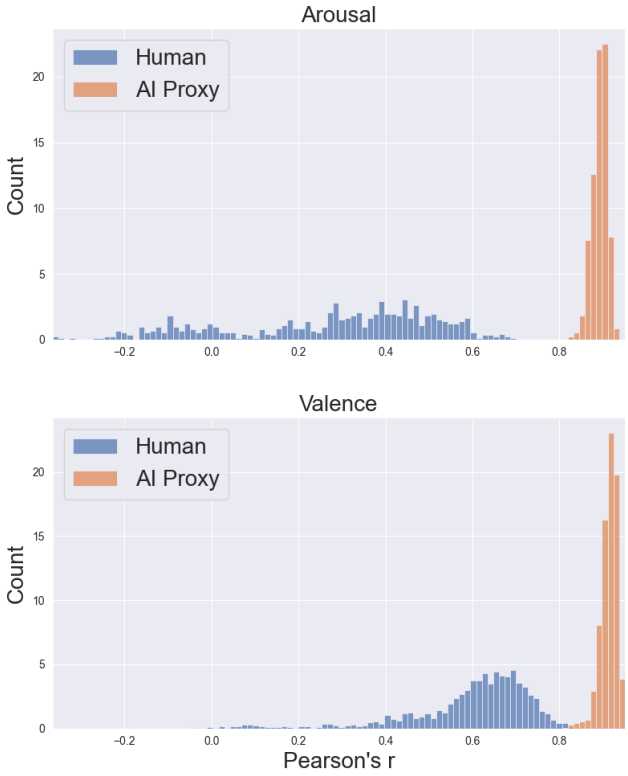

Figure 3: Pairwise inter-rater alignment among human and AI proxy raters for arousal (top) and valence (bottom).

Acknowledgment

Feedback from GPT-4o was used to assist in the writing and editing of this report.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., ... Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS nexus*, *3*(7), pgae245.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., & Finn, C. (2024). *Persona: A reproducible testbed for pluralistic alignment.* Retrieved from https://arxiv.org/abs/2407.17387

Chang, E. Y. (2024). *Modeling emotions and ethics with large language models.* Retrieved from https://arxiv.org/abs/2404.13071

eun Yoon, S., He, Z., Echterhoff, J. M., & McAuley, J. (2024). *Evaluating large language models as generative user simulators for conversational recommendation.* Retrieved from https://arxiv.org/abs/2403.09738

He, Z., Guo, S., Rao, A., & Lerman, K. (2024). *Whose emotions and moral sentiments do language models reflect?* Retrieved from https://arxiv.org/abs/2402.11114

Jones, C. R., & Bergen, B. K. (2024). *People cannot distinguish gpt-4 from a human in a turing test.* Retrieved from https://arxiv.org/abs/2405.08007

Jones, C. R., & Bergen, B. K. (2025). *Large language models pass the turing test.* Retrieved from https://arxiv.org/abs/2503.23674

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, *15*(4), 241–252.

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (oasis). *Behavior research methods*, *49*, 457–470.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, *14*(1), 21445.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). *Large language models: A survey.* Retrieved from https://arxiv.org/abs/2402.06196

Ogg, M., Bose, R., Scharf, J., Ratto, C., & Wolmetz, M. (2025). *Turing representational similarity analysis (rsa): A flexible method for measuring alignment between human and artificial intelligence.* Retrieved from https://arxiv.org/abs/2412.00577

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., ... Malkov, Y. (2024). *Gpt-4o system card.* Retrieved from https://arxiv.org/abs/2410.21276

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature reviews neuroscience*, *9*(2), 148–158.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sabour, S., Liu, S., Zhang, Z., Liu, J. M., Zhou, J., Sunaryo, A. S., ... Huang, M. (2024). *Emobench: Evaluating the emotional intelligence of large language models.* Retrieved from https://arxiv.org/abs/2402.12071

Tiganj, Z., Dickson, B., Maini, S. S., & Nosofsky, R. (2025, Jan). *Comparing perceptual judgments in large multimodal models and humans.* PsyArXiv. Retrieved from osf.io/preprints/psyarxiv/pcmrj_v2 doi: 10.31234/osf.io/pcmrj_v2

Zhao, S. C., Hu, Y., Lee, J., Bender, A., Mazumdar, T., Wallace, M., & Tovar, D. A. (2025). *Shifting attention to you: Personalized brain-inspired ai models.* Retrieved from https://arxiv.org/abs/2502.04658

Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2025). *Large language models as proxies for theories of human linguistic cognition.* Retrieved from https://arxiv.org/abs/2502.07687