Measuring Alignment between Human and Artificial Intelligence with Representational Similarity Analysis

Mattson Ogg (mattson.ogg@jhuapl.edu) and Michael Wolmetz (michael.wolmetz@jhuapl.edu)
Research and Exploratory Development Department
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723

Abstract:

Large Language Models (LLMs) are improving at an incredible rate. With increasing scale comes emergent properties, including an ostensibly human-like understanding of the world. However, it is difficult to assess how these models process and represent information and it is not clear how best to measure their similarities with humans. To help meet this need, we developed a generalizable behavioral task for LLMs (sometimes called a "Turing Experiment") based around pairwise behavioral ratings to facilitate a representational similarity analysis (RSA) that measures alignment among LLM and human agents. Using this method, which we refer to as "Turing RSA," we quantified how aligned the similarity ratings that different LLMs provided for a well-studied set of stimuli from the cognitive neuroscience literature were to human responses at a group and individual level. We found GPT-4 to be the best current proxy of human behavior among its family of models across text and image modalities, but that the inter-individual variability among human participants is hard to reproduce with LLMs. We show that RSA helps us understand how LLMs encode knowledge about the world, examine the variability among agents, and measure their representational alignment with humans.

Keywords: Large Language Models, Explainability, Representational Similarity, Artificial Intelligence

Introduction

The rapid pace of LLM improvement (Kaplan et al., 2020) requires a scalable, generalizable method for understanding how these powerful models represent knowledge about the world and how these representations might differ from human knowledge. This is complicated by the ostensibly black box nature of LLM reasoning. However, this opacity resembles the problem the human mind poses for cognitive neuroscientists. We therefore expect that tools developed in that field (and related ones) will be of great use in understanding these large neural networks.

One of the most powerful methods for mapping the structure of how a human participant represents information about the world, is the use of similarity or dissimilarity ratings with respect to a pair of stimuli (Shepard, 1980). This straight forward but high-level task is adaptable to a wide array of domains and questions (e.g., "How similar are the words 'apple' and 'hand?'" or "How similar are these two images?"). Ratings given by participants on each trial comprise a pairwise behavioral distance metric (Hout et al., 2013). This task is especially useful when the experimenter does not have direct access to the participant's internal representations (e.g., neural responses or model embeddings), such as for LLMs. Representational similarity analysis (RSA; Kriegeskorte & Kievit, 2013) or representational alignment (Sucholutsky, et al. 2023) is an analysis framework that leverages correlations among such distance, or dissimilarity, matrices ("DSMs"), thereby quantifying the alignment of different representational spaces. This method has been applied to understand representational geometries across diverse systems including different organisms (Kriegeskorte et al., 2008), individuals, brain regions (Cichy et al., 2014; Giordano et al., 2023; Ogg et al., 2019), behavior (Carlson et al., 2014; Ogg & Slevc, 2019) or computational models (e.g., Meher et al., 2021; Ogg & Skerritt-Davis 2021).

We designed a "Turing Experiment" (Aher et al., 2023; Mei et al., 2024) we term "Turing RSA" that adapted a pairwise similarity rating task from the cognitive neuroscience literature to probe the representational geometry of frontier LLMs via RSA. We asked different simulated LLM participants to provide a similarity rating for a set of well-studied text and image stimuli (but see Grootswagers & Robinson, 2021) and used these ratings to quantify the alignment of knowledge representations between (and within) groups of LLM and human participants.

Methods

We elicited responses from different versions of Open AI's Generative Predictive Transformer (GPT) models (Brown et al., 2020) via the Open AI Azure API. We used a GPT-3.5 Turbo model ("gpt-35-turbo-16k," version "0613"), a GPT-4 model (version "1106-preview"), and a GPT-4-vision model (version "vision-preview"). For each experimental run (i.e., simulated

participant) we initialized the model, using a temperature of 1.0 (higher than the default 0.7, to encourage diverse responses). Participants were simulated following Aher and colleagues (2023) by assigning the LLM agent a surname (Garcia, Jeanbaptiste, Kim, Nguyen, Olson, Rodriguez, Smalls, Snyder) and honorific (Ms., Mr., Dr.). The cost of the vision models was higher so these were run with just four surnames and without the "Dr." honorific.

On each trial the model was asked to rate how similar a pair of 67 word (from Carlson et al., 2014) or image (from Kriegeskorte et al., 2008; Cichy et al., 2016) stimuli were on a scale from 0 to 100. We compared LLM ratings with: 1) human semantic relatedness ratings for the class label words (from Carlson et al., 2014) and 2) with behavioral (SPOSE) embeddings learned to predict a large number of leave-one-out behavioral ratings for images from the THINGS dataset (see Hebart et al., 2020). Comparisons with SPOSE embeddings were restricted to the 55 overlapping object classes between the stimulus sets.

We averaged the responses among participants for each model (or among humans) and computed a Spearman rank correlation among the flattened, group-level dissimilarity matrices. To evaluate inter-subject agreement, we computed correlations among pairs of individual LLM or human participant's DSMs.

Results

The results of this experiment are conveyed in Figure 1, which shows the group-averaged DSMs and the Spearman correlations among them (all $p < 0.05$). GPT-4 ratings were the most similar to the human participant's text ratings (r_s = 0.70), showing much better alignment than GPT-3.5 (r_s = 0.46). The GPT-4-Vision alignment with human data was also modest. Notably, GPT-4 text-only ratings were better aligned with human SPOSE embeddings which were generated based on images rather than text. Similar results were obtained when GPT-4-Vision rated pairs of images from the THINGS database.

Model variability is summarized in Figure 2, showing that the increased alignment of GPT-4 models comes at a cost of greater homogeneity among individual participant ratings that thus fail to capture the natural variability of human participants (despite a higher temperature value). Indeed, no model reproduced both the overall range and median inter-participant alignment observed among the human participants.

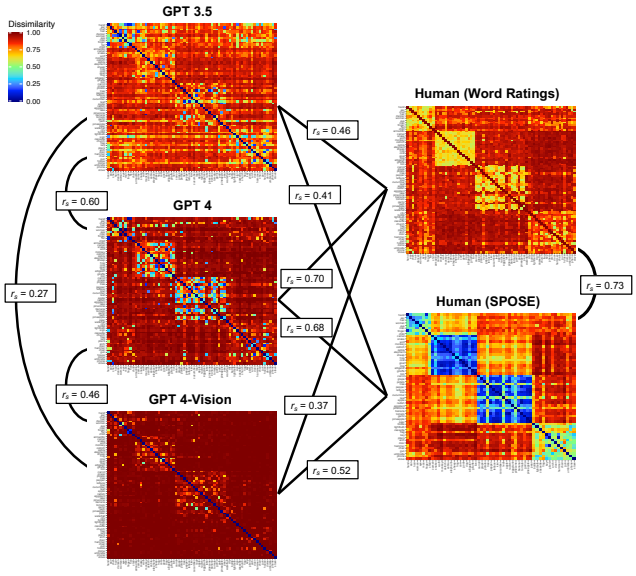

Figure 1: DSMs for each model system averaged over individual participants (except SPOSE) along with the Spearman rank correlations among them (all $p < 0.05$).

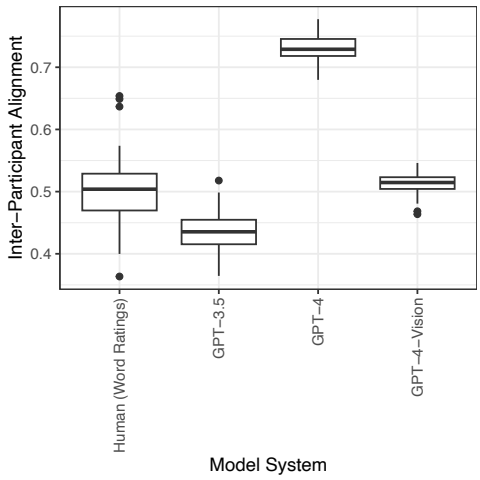

Figure 2: Inter-participant variability of human and LLM participants via Spearman rank correlation among each pair of participants for each model (y-axis).

Discussion

We report a powerful, flexible Turing Experiment for querying LLM knowledge and measuring alignment with human representations, demonstrating that GPT-4's text representations capture semantic similarity across input domains and align most closely with human judgments. However, inter-individual variability is still an outstanding issue among the LLMs we evaluated, as no model adequately reflected this dimension of human behavior.

Acknowledgments

We acknowledge support from the Independent Research and Development (IRAD) Fund from the Research and Exploratory Development Mission Area. We thank Ritwik Bose, James Scharf and Christopher Ratto for helpful discussions in the course of this work and Robert Slevc for collecting and sharing the human behavioral data.

References

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*(4468), 390-398.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(1), 93-103.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401-412.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... & Griffiths, T. L. (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience, 26*(4), 664-672.

Ogg, M., Carlson, T. A., & Slevc, L. R. (2020). The rapid emergence of auditory object representations in cortex reflect central acoustic attributes. *Journal of Cognitive Neuroscience, 32*(1), 111-123.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience, 17*(3), 455-462.

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2014). The emergence of semantic meaning in the ventral temporal pathway. *Journal of Cognitive Neuroscience, 26*(1), 120-131.

Ogg, M., & Slevc, L. R. (2019). Acoustic correlates of auditory object and event perception: Speakers, musical timbres, and environmental sounds. *Frontiers in Psychology, 10*, 450659.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences, 118*(8), e2011417118.

Ogg, M., & Skerritt-Davis, B. (2021). Acoustic Event Detection Using Speaker Recognition Techniques: Model Optimization and Explainable Features. In *DCASE* (pp. 80-84).

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023, July). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning* (pp. 337-371). PMLR.

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.

Grootswagers, T., & Robinson, A. K. (2021). Overfitting the literature to one set of stimuli and data. *Frontiers in Human Neuroscience, 15*, 682661.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex, 26*(8), 3563-3579.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*(11), 1173-1185.