Extracting Object Feature Norms from Large Language Models

Stephen Mazurchuk (smazurchuk@mcw.edu)
Biophysics, Medical College of Wisconsin, 8701 Watertown Plank Rd.
Milwaukee, WI 53226 United States

Andrew Anderson (andanderson@mcw.edu)
Neurology, Medical College of Wisconsin, 8701 Watertown Plank Rd.
Milwaukee, WI 53226 United States

Abstract:

Many cognitive neuroscience researchers collect human semantic feature ratings for words. While these ratings are often collected using crowdsourced systems, it can be costly, and a considerable amount of work can be required to prepare the query. Consequently, there is growing interest in augmenting this task by using the representations contained in large language models. We present and validate a method for extracting accurate semantic ratings using a simple pairwise comparison paradigm that allows researchers to easily estimate human norms using freely available open-source language models.

Keywords: Large Language Model; Semantic; Feature

Introduction

Collecting human norms is a major task in neuroscience research. One question that is receiving increased attention is whether large language models (LLMs) can help facilitate the collection of normative data. One way to approach this task is to directly query LLMs with the same prompts provided to human raters. One author has taken this approach and shown success in recapitulating human norms (Trott, 2024). While this approach is viable, it has some limitations which stem from using GPT-4. Namely, the cost associated with using a proprietary model, discretized ratings, and no clear approach for assessing model confidence. We provide an alternative approach that mitigates these concerns as well as provides rich information that allows for estimation of confidence in human rating.

The work presented here is largely derivative of the paper and corresponding data provided by Grand, Blank, Pereira, and Fedorenko (2022). Our manuscript uses the human ratings collected from that study and uses their work with GloVe to approximate ratings (Pennington, Socher, & Manning) as a benchmark. Their approach, called semantic projection, estimates ratings for a given feature by creating a difference vector from adjectives describing that feature (e.g., Size: Large – Small). Word vectors are then projected along this vector to estimate human ratings. The purpose of our manuscript is to validate a simple approach for extracting human ratings from large language models (LLMs).

Previous Approaches

Although directly querying LLMs proved effective (Trott 2024), that approach is much less reliable with smaller LLMs and requires the use of closed-source proprietary models only accessible through the internet. A further downside to direct querying is that since the queries typically use Likert scales, the outputs are discretized ratings. Our approach provides continuous ratings. Further, a measure of confidence is possible.

Another approach for estimating human norms from distributional models is to fit a linear regression transformation from a large distributional space to a semantic space (Wang et al., 2023). The primary downside of this approach is that one needs to already have at least some human ratings with which to determine the transformation that relates the two spaces. This does not allow for new features for which there is no human data to be estimated. The approach of Grand et al. (2022) allows for new features to be estimated for which ratings are not already collected, and that is why we chose to benchmark relative to this approach.

Methods

The basic premise of our approach is shown in **Figure 1**. The task of ordering a set of stimuli along some semantic feature is accomplished by generating prompts for all pair-wise comparisons by a given semantic feature. While half of these comparisons appear redundant, it is empirically observed that the

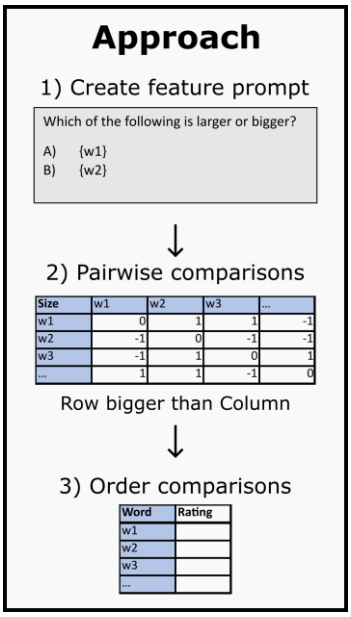

Figure 1: Illustration of pairwise comparison approach. Each query is a pairwise comparison that fills out the entries of a square comparison matrix. The comparison matrix can be converted to ratings by simply summing across rows or columns.

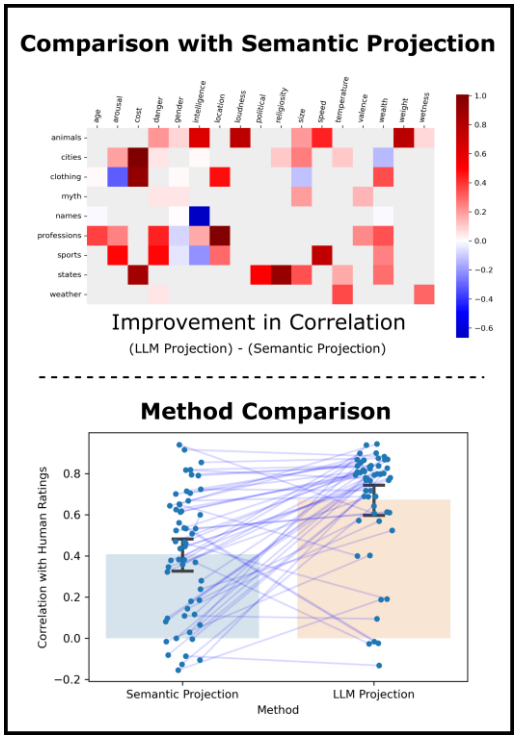

Figure 2: Top: This plot shows the difference in correlations between LLM projection and human ratings compared to Semantic projection with human ratings. **Bottom**: The difference in correlation is visualized demonstrating a significant improvement.

order of the options can have effect. We recapitulated the work of Grand et al. (2022) by converting all features into a series of forced-choice prompts. These prompts were then evaluated on the Hugging Face implementation of the Google Gemma-7B-Instruct model. The outputs were then processed into a pair-wise comparison matrix. This matrix was converted into ratings by summing across the rows and then subtracting the sum across the columns. Human rater agreement was assessed using a single-measures consistency based intraclass correlation coefficient (ICC). A python script to generate these results will be made available on GitHub.

Results

Shown in **Figure 2**, We found that querying the LLM with a forced choice prompt improved the correlation between the predicted rating and human rating relative to correlations found using semantic projection (mean r = .26, Wilcoxon p < .001). As observed in **Figure 3**, we also observed that the intraclass correlation (ICC) between human raters was predictive of the correlation between the LLM and average of human ratings. Lastly, as shown in **Figure 3**, we observed that the standard

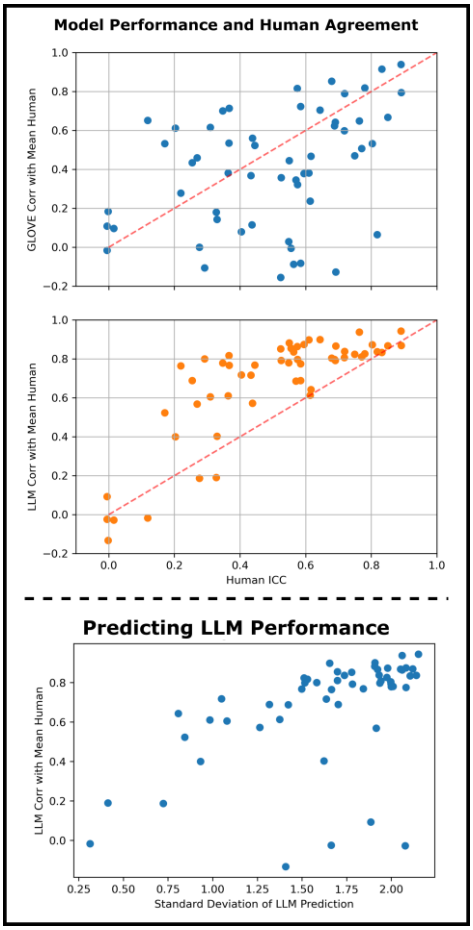

Figure 3: Top: Human rater agreement is a good predictor of which features semantic projection will work well for. **Bottom**: Standard deviation of LLM-predicted vector predicts how well the vector will correlate with human ratings. For all plots, each point represents a feature.

deviation of the LLM predicted rating vector was predictive of how well the LLM rating vector correlated with the mean human rating.

Conclusions

We found that querying LLMs with forced-choice prompts can be an effective way to extract semantic feature representations. This approach is more likely to be effective when human raters are likely to agree, and the correlation between the LLM prediction is somewhat predicted by the standard deviation of the LLM prediction. Future work involves better estimation of the model performance. For example, the number of transitive violations in the pairwise comparison matrix could be a marker of how well a feature is likely to correlate with a human rating.

Acknowledgments

We would like to thank the authors of Grand et al. (2022) for making their data available.

References

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour, 6*(7), 975-987. doi:10.1038/s41562-022-01316-8

Pennington, J., Socher, R., & Manning, C. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. doi:10.3115/v1/d14-1162

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, 1-19. doi:10.3758/s13428-024-02337-z

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data, 10*(1), 106. doi:10.1038/s41597-023-01995-6