AI-ENHANCED SEMANTIC FEATURE NORMS FOR 786 CONCEPTS

Siddharth Suresh*[1,2,6], **Kushin Mukherjee***[1,2], **Tyler Giallanza**[3,4], **Xizheng Yu**[5], **Mia Patil**[2,6], **Jonathan D. Cohen**[3,4], and **Timothy T. Rogers**[1,2]
[1]Department of Psychology, University of Wisconsin-Madison, [2]Wisconsin Institute for Discovery,
[3]Princeton Neuroscience Institute, [4]Department of Psychology, Princeton University,
[5]Department of Computer Science, Brown University,
[6]Department of Computer Science, University of Wisconsin-Madison
*denotes equal contribution, correspondence to

ABSTRACT

Semantic feature norms have been foundational in the study of human conceptual knowledge, yet traditional methods face trade-offs between concept/feature coverage and verifiability of quality due to the labor-intensive nature of norming studies. Here, we introduce a novel approach that augments a dataset of human-generated feature norms with responses from large language models (LLMs) while verifying the quality of norms against reliable human judgments. We find that our AI-enhanced feature norm dataset shows much higher feature density and overlap among concepts while outperforming a comparable human-only norm dataset and word-embedding models in predicting people's semantic similarity judgments. Taken together, we demonstrate that human conceptual knowledge is richer than captured in previous norm datasets and show that, with proper validation, LLMs can serve as powerful tools for cognitive science research.

Keywords: semantic knowledge; feature listing; large language models; similarity judgments

1 INTRODUCTION

The study of human conceptual knowledge has relied on semantic feature norms — representations of concepts in terms of their associated features — since their introduction by Rosch in the 1970s (Rosch, 1975). Norming studies present participants with a set of concepts and, for each, asks them to list as many characteristic properties as they can. Aggregating features across items and participants creates semantic vectors the elements of which correspond to the elicited features and the entries of which indicate whether people regularly judge the concept to possess the corresponding property. Proximity between two such feature vectors relates systematically to their perceived semantic relatedness—thus lions and tigers are viewed as similar kinds of things because they have many overlapping and fewer distinguishing properties. Norming datasets collected over the years (McRae et al., 2005; Devereux et al., 2014; Buchanan et al., 2019; Ruts et al., 2004; Hansen & Hebart, 2022; Dilkina et al., 2008) have helped to answer questions about the organization of semantic memory (Collins & Loftus, 1975; Ashcraft, 1978), its degradation in semantic disorders (Farah & McClelland, 2013; Rogers & McClelland, 2004; Garrard et al., 2001; Cree & McRae, 2003), its relationship to control (Giallanza et al., 2024), and its neural bases (Cox et al., 2024; Clarke & Tyler, 2014) (see Kumar (2021) for a review).

Semantic norming requires extensive human labor both in data collection and curation/post-processing. Prior studies have met this challenge in different ways, each requiring some degree of compromise as elaborated below. Other recent work has sought alternatives to human feature norms by making use of natural language processing technologies, including word embeddings from methods such as word2vec and GloVe (Pennington et al., 2014; Mikolov et al., 2013) as well as feature norms generated artificially by large language models (LLMs) (Hansen & Hebart, 2022). However, word embeddings fail to capture the semantic structure perceived by humans as effectively as feature norms, and their dimensions lack the transparent interpretability of feature-based representations, at

least for for concrete objects (Suresh et al., 2023b;a). LLMs can generate super-human lists of features that go far beyond what a typical person might know, however they also frequently confabulate properties that are untrue -— the well documented 'hallucination problem' (Huang et al., 2024).

LLMs have demonstrated remarkable alignment with human behavior across diverse tasks, ranging from semantic similarity judgments to higher-order reasoning (Hagendorff, 2023; Street et al., 2024; Giallanza & Campbell, 2024; Dasgupta et al., 2022; Kosinski, 2024; Binz & Schulz, 2023; Chuang et al., 2023b;a; Mukherjee et al., 2024; Sucholutsky et al., 2023). Although these models are prone to hallucination (Bender et al., 2021; Ji et al., 2023; Farquhar et al., 2024; Xu et al., 2024) —occasionally generating inaccurate or spurious outputs—they have been successfully harnessed to create synthetic datasets and experimental stimuli for both cognitive science and ML applications (Trott, 2024; Hansen & Hebart, 2022; Gupta et al., 2023; Wu et al., 2024; Patel, 2024).Building on these insights, our work leverages LLMs to construct a large-scale feature norming dataset that synergistically integrates robust human judgments with machine-generated enhancements, thereby bridging the gap between human fidelity and computational scalability.

The current work seeks a middle way between human-only and machine-only norm generation. We crowd-sourced feature lists for a modestly large and representative set of 786 concrete object concepts thus ensuring that the features included in the set are those that human participants discern. We then used LLMs to aid in the most labor-intensive parts of data curation and post-processing, resulting in a novel *AI-enhanced* set of semantic feature norms. We illustrate remarkable differences between human-only and AI-enhanced norm sets, then report empirical studies designed to assess whether the AI-enhanced norms capture human-perceived semantic structure better than do human-only norms or "out-of-the-box" word embeddings.

2 Related work

Human-Centric Semantic Norms The use of semantic feature norms has a rich history in cognitive science, beginning with early work by Rosch (1975) and further developed in datasets such as those by McRae et al. (2005), Devereux et al. (2014), and others. (Buchanan et al., 2019; Ruts et al., 2004; Hansen & Hebart, 2022; Dilkina et al., 2008). These datasets have been instrumental in exploring the organization of semantic memory(Collins & Loftus, 1975; Ashcraft, 1978), its degradation in clinical populations(Farah & McClelland, 2013; Rogers & McClelland, 2004; Garrard et al., 2001), and even the neural underpinnings of concept representation(Cox et al., 2024; Clarke & Tyler, 2014). Their transparent, interpretable structure has also made them useful in understanding the cognitive basis of semantic similarity.

Machine Learning and NLP Representations Parallel to human-based approaches, the machine learning community has advanced distributed semantic representations through methods like word2vec(Mikolov et al., 2013) and GloVe(Pennington et al., 2014). Although these models enable large-scale applications, their latent dimensions often lack clear semantic meaning when compared to feature norm-based representations. More recently, LLMs such as GPT-3 and BERT(Brown et al., 2020; Devlin et al., 2018) have been tapped to generate(Hansen & Hebart, 2022) and verify semantic features (Suresh et al., 2023a). However, challenges remain regarding the factual accuracy and reliability of these models due to issues such as hallucination(Huang et al., 2024).

Cogntive Science and LLMs Recent work in cognitive science has increasingly leveraged LLMs both as experimental subjects and as computational models of human cognition. For instance, studies such as Misra et al. (2022); Marjieh et al. (2022) have demonstrated that LLMs can simulate human-like responses in psycholinguistic tasks, while Marjieh et al. (2023) highlights their use in generating stimuli that capture subtle semantic variations. Other research (Suresh et al., 2023b) has compared the internal representations of LLMs with human behavioral data, revealing notable parallels and differences in semantic memory organization. Furthermore, studies such as Campbell et al. (2024), Binz et al. (2024), and Marjieh et al. (2024) have tried understanding LLMs and VLMs using behavioral tasks grounded in Cognitive Science. Collectively, these works underscore the dual role of LLMs as both scalable experimental proxies and as computational frameworks for understanding human cognition.

Hybrid Human–AI Systems and Interpretability Our work aligns with emerging trends in hybrid human–AI systems for data annotation and curation(Trott, 2024; Hansen & Hebart, 2022; Gupta

et al., 2023; Wu et al., 2024; Patel, 2024). By combining the scalability of LLMs with the reliability of human judgments, we aim to overcome the limitations inherent in both approaches. Moreover, our AI-enhanced feature norms offer a pathway to help out both Cognitive Scientists and ML Interpretibility researchers. Feature norm semantic dimensions that can be used for explainable AI, thereby bridging gaps between human cognition and machine learning.

3 Study I: Building an LLM-assisted semantic norm dataset

3.1 Overview of the approach

Human feature-norming studies involve up to 4 steps, each requiring significant effort and thus are subject to constraints that can limit the resulting data. Here we consider each step, limitations faced by prior studies, and the approach taken in the current work. The overall workflow for our approach is shown in Figure 1.

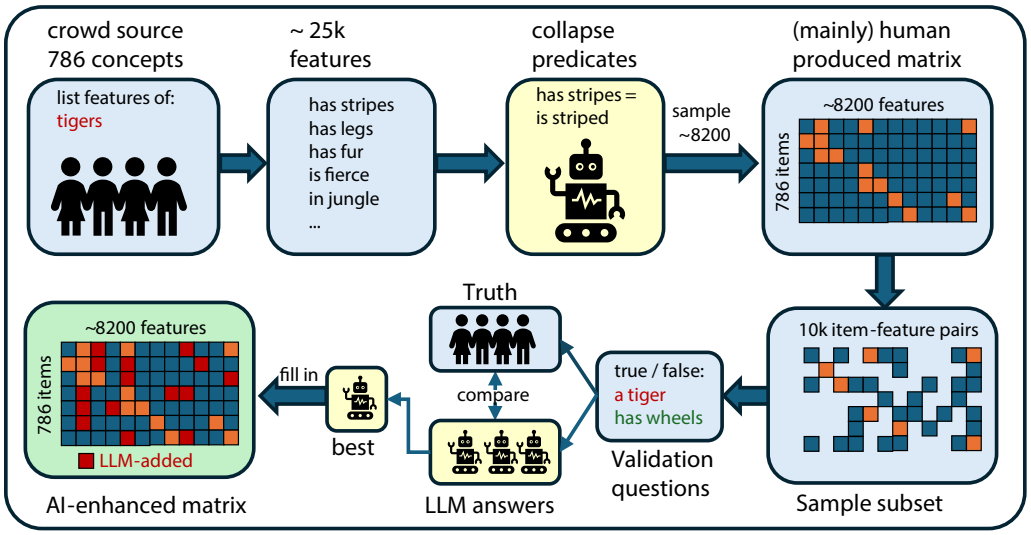

Figure 1: A schematic representation of our workflow. Features were initially crowd-sourced for 786 concepts, forming a human-generated matrix. A subset of 10,000 concept-feature pairs underwent validation via human judgments. LLM responses were compared to these human judgments to determine the best-performing strategy. Using this method, LLMs completed the matrix for all 8,200 selected features, forming the AI-augmented matrix.

Concept selection. The structure appearing in a given dataset depends on the concepts included. Early norms, used in semantic memory studies, focused on hierarchically structured, easily nameable concepts (e.g., animals, plants), often excluding typical examples (e.g., robins, sparrows) in favor of atypical ones (e.g., penguins, ostriches) and omitting concepts that don't fit neatly into these hierarchies. To improve representativeness, we included all 565 concepts from the Ecoset dataset Mehrer et al. (2021), which comprises frequent, unambiguous basic-level concrete object names, along with items from the McRae (McRae et al., 2005) and Leuven (De Deyne et al., 2008) norms. We also added superordinate categories (e.g., animal, vehicle) and higher-frequency subordinate names (e.g., robin, trout) to better capture domain substructure. The final set comprised **786** concrete object concepts.

Feature elicitation. In common with other recent norming studies (De Deyne et al., 2008; McRae et al., 2005; Ruts et al., 2004; Devereux et al., 2014), we elicited features from participants on Amazon Mechanical Turk using procedures described below.

Feature reduction. Norming studies typically yield a large set of unique features, most appearing in a single concept. To manage this complexity, researchers often consolidate distinct yet semantically related properties – e.g., if *is hairy* and *is furry* are used by different participants to describe a' coconut', these features may be deemed as equivalent and thus a single feature allowing for fea-

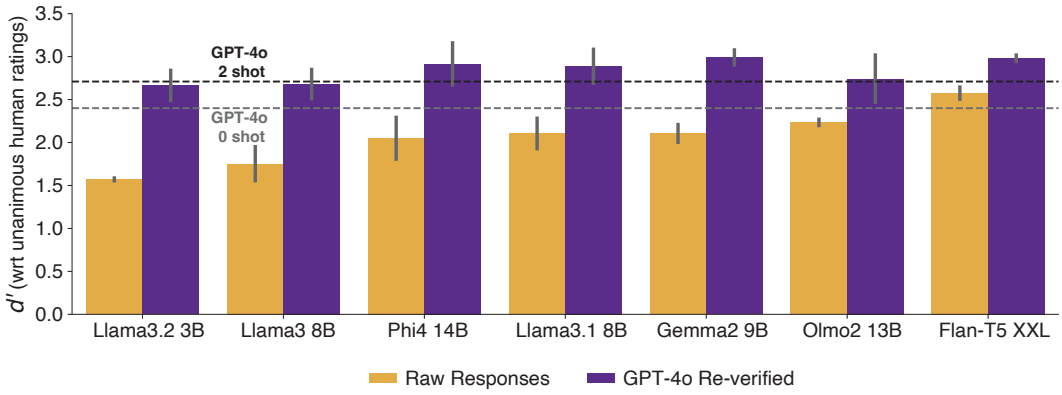

Figure 2: Models' ability to reliably predict human feature-concept ratings measured using d' using raw responses (orange) and responses re-verified using GPT-4o. Bar heights are mean d' across the 0-shot and 2-shot experiments. Gray and black dashed lines correspond to GPT-4o's performance in the 0-shot and 2-shot setting respectively. Errorbars correspond to bootstrapped 95% confidence intervals.

ture overlap with concepts possessing both *is hairy* (e.g., ape) and *is furry* (e.g., rabbit). While this process simplifies the feature space and enhances conceptual similarity across items, it is labor-intensive and relies on subjective human judgments. Here, we extracted using phrase embeddings from `GPT-3` to perform minimal feature collapse. Specifically, we extracted embeddings for featural descriptions (e.g., *has a furry outer layer*) and merged only highly similar clusters, merging propositions with near-identical semantic content but variable wording (e.g., *has a furry outer layer ',is furry*, and *feels furry)* while still distinguishing them from close synonyms (e.g., *is hairy*). This automated approach reduced the initial 25k raw features to approximately ∼20k features, from which we randomly sampled ∼8,200 features for subsequent analysis.

Feature verification. The features that participants generate in the elicitation phase typically constitute a fraction of what they actually know. For this reason, some norming studies conduct a *feature verification* step where human participants consider every concept/feature pair and judge whether the feature is true of the concept (De Deyne et al., 2008; Dilkina et al., 2008). This step greatly enriches the structure encoded in the norms. For instance, most participants list the feature *has a long neck* for giraffes and swans but for few other items. Yet when asked, most participants agree that *has a long neck* is true of items as varied as a duck, a beer bottle, and a cello. Thus, the verification phase surfaces people's latent knowledge that they don't generate spontaneously. Since the number of concept/feature pairs grows exponentially, this is by far the most labor-intensive part of the process and prior studies have either employed a relatively modest set of concepts and features (Dilkina et al., 2008) or have limited verification only to specific semantic domains (De Deyne et al., 2008). We leveraged LLMs to conduct the feature-verification phase – first comparing different models and strategies in their ability to capture human judgments on a set of 6,122 concept-feature pairs where human participants showed unanimous agreement, then using the most successful strategy to verify all ∼6.5M concept/feature pairs, producing an AI-enhanced norm set.

3.2 METHODS

Human feature elicitation. This phase provided human-elicited data for all concepts in the set, providing the raw features from which human-only and AI-enhanced norms were derived.

Participants. 50 participants were recruited through Amazon Mechanical Turk and were compensated 4$ for the task which would require 20 minutes to complete. The study was approved by the home university's IRB.

Stimuli and procedure. Stimuli were 786 concrete object nouns. Using a web-based interface, each participant viewed up to 75 different words in randomized order, and for each typed in as many different features as they could generate. The instructions emphasized generating various types of features, including physical/perceptual features (appearance, smell), functional features (uses,

contexts), and other characteristics. Participants were asked to format their responses as individual features per line using standardized phrasing (e.g., "has ears" rather than "a dog is an animal that has ears").

3.2.1 HUMAN FEATURE VERIFICATION.

This phase had human participants verify ∼10k concept-feature pairs, providing an empirical basis for evaluating the performance of different AI-aided approaches to feature verification.

Participants. 556 participants were recruited through Amazon Mechanical Turk and compensated $1.40 for a 5-8 minute task. Participants were allowed to complete multiple sessions contingent upon maintaining satisfactory performance.

Stimuli and procedure. The stimuli were concept-property pairs sampled randomly from results of the feature-elicitation task. Data were collected through an online interface. Each trial paired one concept (e.g. "alligator") with one feature randomly sampled from the full set. The sampled feature could come from any domain or item–for alligator, it could be something reasonable (e.g. "has legs"), something clearly false (e.g. "has wheels") or something uncertain (e.g. "has ears"). For each pair participants judged whether the property is true of the item by pressing a keyboard button. The instructions emphasized that subjective properties should be evaluated based on common consensus (e.g., "cute" for "dog"), and properties that were sometimes true should be marked as true (e.g., "brown" for "dog"). Each participant made about 110 judgments, and we collected 5 or more judgments on each of 10,545 unique pairs. Participants could skip unfamiliar concepts or non-

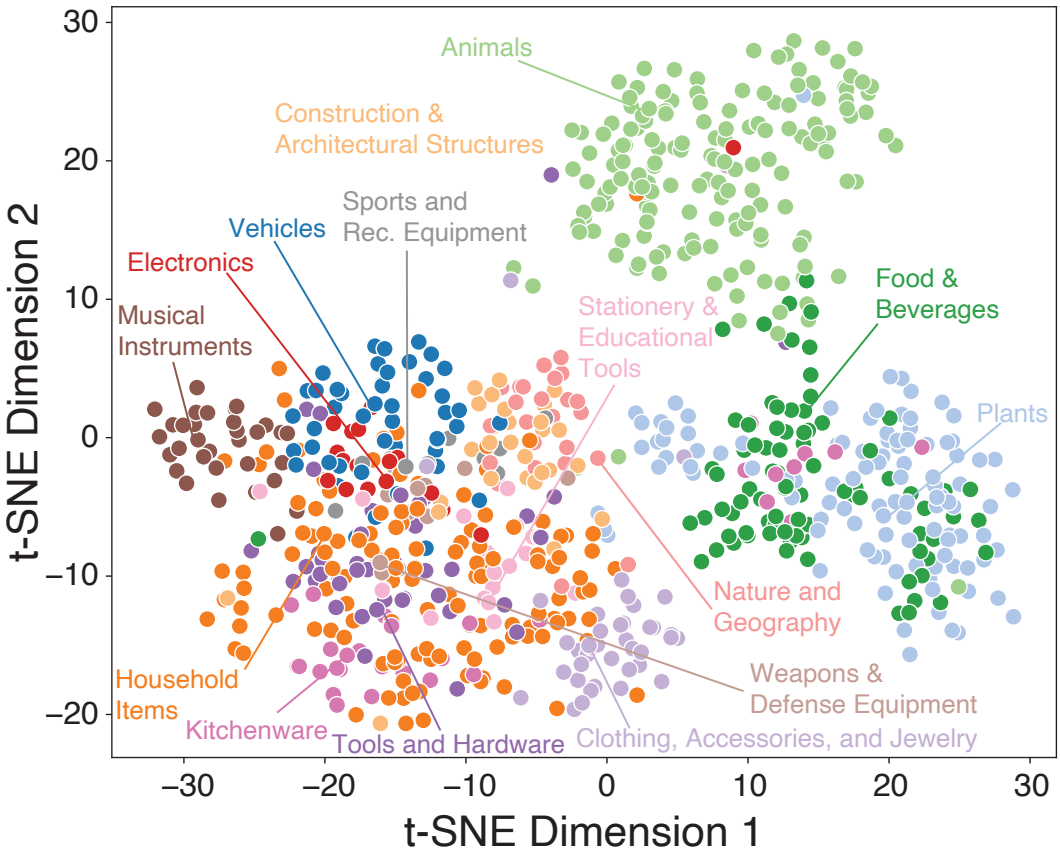

Figure 3: t-stochastic neighbor embeddings of the semantic vectors for each of 787 concepts derived from the final verified matrix. Category labels were generated by combining higher order labels from existing norm datasets and LLM-suggested categories from GPT-4o.

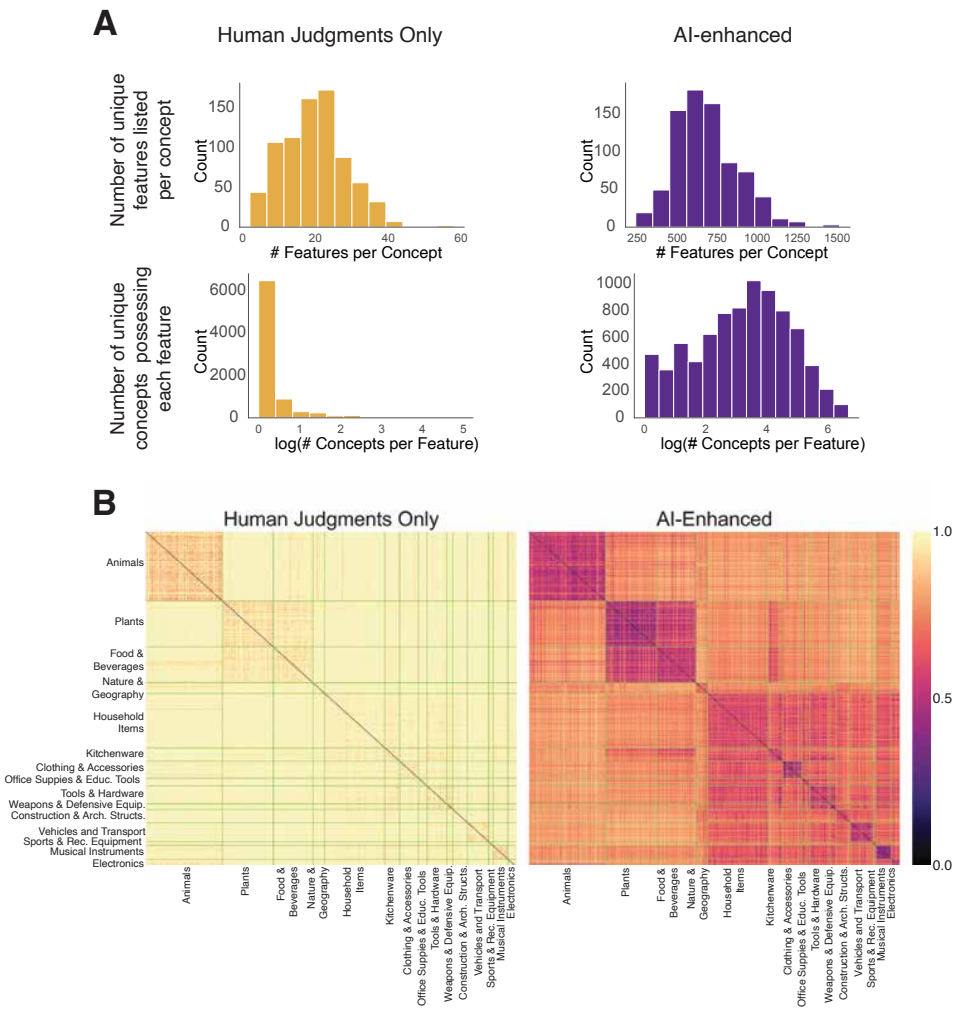

Figure 4: (A) Counts of valid features per concept and number of concepts that share common features for the reduced human-generated matrix (top row) and AI-enhanced norm matrix (bottom row). (B) Pairwise cosine dissimilarity matrices based on the reduced human-generated norm matrix (left) and AI-enhanced norm matrix (right).

sensical properties by pressing the space bar, with skipped items replaced to maintain the required number of judgments.

Machine assisted feature verification.

Our ultimate goal was to use LLMs to complete the feature-verification step for all possible concept/property pairs. Since there are millions of possible pairs, we first considered how well each of several different models and prompting strategies could capture real human judgments on the items collected in the human feature-verification study. In these data participants showed different opinions for about 40% of the items–thus either opinion expressed by an LLM would agree with at least one human participant for these items. We therefore selected the 6,122 concept-feature pairs for which all participants made the same decision (either all yes or all no), and used these decisions as a ground-truth for evaluating LLM performance.

Model Suite. We primarily focused on performant open-sourced models because these are accessible to other researchers for replication purposes and generally more affordable to run. We included open-sourced models that have open weights, are generally high-scoring on standard LLM benchmarks (Hendrycks et al., 2020), and can be run on consumer-grade hardware. Specifically, we

evaluated 3 models from Meta's `Llama` family (`Llama3`, `Llama3.1`, and `Llama3.2`) (Dubey et al., 2024), Microsoft's `Phi-4` (Abdin et al., 2024), Ai2's `Olmo2` (OLMo et al., 2024), and Google's `Gemma2` (Team et al., 2024) and `Flan-T5` (Wei et al., 2021). We evaluated all models at full `bfloat16` precision on a commercially available Nvidia H100 GPU. For comparison to a state-of-the-art closed model, we also evaluated GPT-4o via its API.

Evaluation Protocol. We prompted all models using the following general prompt -

```
In one word True or False, answer the following
question question: Is the property [x] true for
[y]? Answer:
```

...where x was a feature and y was a concept with the square brackets included in the prompt. We ran two prompting experiments: (1) a zero-shot experiment providing the models with just the question above as input, and (2) a two-shot experiment providing the models with two example feature-concept pairs, one true and one false to potentially improve the models' ability to perform the task via in-context learning (Brown et al., 2020). We used the same two examples for all prompts.

Post-processing. To extract meaningful answers from model-generated text we first restricted responses to a maximum of five tokens, then conducted a case-insensitive search of model responses for the strings 'True' or 'Yes' to indicate a positive response, and 'False' or 'No' to indicate a negative response. In rare cases where no match was found we set the model response to 'False'.

3.2.2 Results.

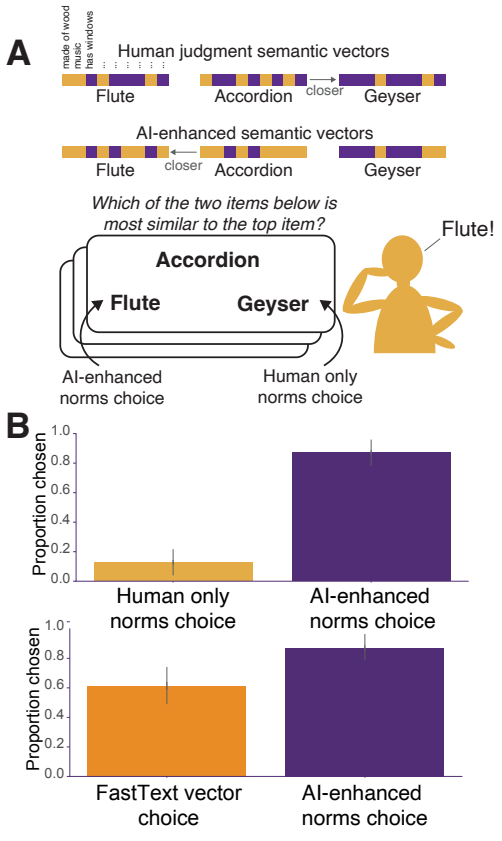

Figure 5: (A) Procedure for generating trials for the triadic judgment experiment and an example trial. (B) Proportion of human responses that aligned with the human matrix (yellow bar) vs. the AI-enhanced matrix (purple bar) and with FastText word embeddings (orange) vs. AI-enhanced semantic vectors (purple) in Experiment 2. Error bars represent standard errors of the means.

To measure how closely LLM responses aligned with unanimous human judgments for the 6,122 feature-concept pairs, we adopted a signal detection approach, treating human responses as the true signal and model responses as guesses. Where humans agreed the property was true of the concept, model guesses were scored as hits if they concurred and misses otherwise. Where humans agreed the property was not true of the concept, model guesses were scored as correct rejections if they concurred and false alarms otherwise. From these counts we computed hit rates and false alarms rates, then converted these to the d' measure of signal discrimination.

The average d' for both zero and two shot conditions can be seen in Figure 2 (yellow bars). Two-shot `GPT-4o` outperformed all open-sourced models, which varied in their match to human responses. Two-shot `Flan-T5 XXL` performed best amongst open-sourced models and better than the zero-shot `GPT-4o`. `Flan-T5's` lower d' relative to `GPT-4o` was driven by a propensity to respond with 'true' to many queries, buoying its hit rate but also increasing its false-positive rate. To preserve the benefits of `GPT-4o` without incurring a prohibitive cost, we next considered a 're-verification' approach in which the 'true' responses generated by a given open-source model were subsequently re-verified by `GPT-4o`, retaining the 'true' value only if both models agreed. The results are shown as purple bars in Figure 2. Re-verification improved performance for all models, surpassing `GPT-4o`

alone. `Flan-T5 XXL` remained a top model, closely matched by `Gemma2 9B`. Given the strong baseline performance of `Flan T5`, we chose this model with `GPT-4o` re-verification to fill out the full semantic feature matrix.

Using Flan T5 and GPT-4o to impute the AI-enhanced matrix. In the human-only matrix, entry $[i, j]$ has a value of 1 wherever a participant produced feature j for concept i and a 0 in all other entries. For every 0 in this matrix, we prompted `Flan T5 XXL` to decide whether the corresponding property is / is not true of the corresponding concept. Where the model decided 'not true,' the zero value was retained in the matrix. Where the model decided 'true,' (534,010 out of 6,436,554 possible pairs) we prompted `GPT-4o` with the same pair to re-verify the answer. If `GPT-4o` agreed the property was true, the cell value was replaced with 1, otherwise the 0 value was retained. This procedure yielded the final **AI-enhanced** norms matrix. Figure 3 shows tsne-based embeddings of all concepts from this matrix.

The AI-enhanced matrix differed remarkably from the human-only matrix in its feature density. While the human matrix has about 20 features per concept on average, the AI-enhanced matrix has about *700* (Figure 4A), and while the majority (78%) of features in the human-only matrix are true of just one concept, this is true of just 5% of features in the AI-enhanced matrix. The increased feature density produces much more richly-structured similarity relations, as shown by the heat plot of pairwise distances between concepts in Figure 4B. While some of this difference may be attributable to false-positives in the AI-enhanced dataset, the comparison to human judgments suggests that the LLM verification strategy is quite good at discriminating true positives from true negatives ($d' > 3.0$). Thus the result suggests that human knowledge about features of concepts may be considerably richer than prior norming studies have suggested.

4 Study 2: Using the new norms dataset to predict human semantic judgments

To assess whether the AI-enhanced norms capture information about semantic structure beyond human-only norms or other approaches, we compared different approaches in their ability to predict human behavior in a triadic similarity judgment task. In this task, participants must decide which of two option concepts is semantically more similar to a target concept. A candidate semantic embedding can "predict" human decisions by selecting whichever option word lies closer to the target word in the embedding space. We can assess the quality of the embedding by comparing how often the predicted response agrees with actual human decisions. In this study, we compared the predictions of the AI-enhanced model to predictions based on the human-only feature norms and to those generated by a common word-embedding approach (`FastText`).

We selected triplets designed to maximally discriminate the AI-enhanced and human-only feature norms. Thus for each trial, one of the option items was closer to the target in the human-only space while the other was closer in the AI-enhanced space (see Figure 5). We then computed how often the majority-vote across human participants agreed with the predictions of each embedding (AI-enhanced, human-only, FastText). If the AI-enhanced norms contain information irrelevant to human-perceived semantics, their predictions should agree with human judgments less often than do those of the human-only norms. Furthermore, if either set of norms simply recapitulates the semantic structure evident in word embeddings, then predictions from the norms should do about as well as predictions from the FastText embeddings.

Generating maximally disagreeing triplets. To generate triplets that maximally differentiated the human-only and AI-enhanced norms, we computed cosine dissimilarity matrices for each set (Figure 4B), Procrustes-aligned them to minimize disparity, and identified concepts with the largest discrepancies in their distances to other concepts. This is described in Equation (1) where where d_{ik} represents the distance between concepts i and k [1]. For example, in the AI-enhanced space, 'accordion' was closer to 'flute' than to 'geyser', while the reverse was true in the human-only space (Figure 5A). We constructed 1,424 triplets where the two matrices produced divergent predictions,

[1] A detailed explanation of the discrepancy metric can be found in the Appendix A.1

with each of the 786 concepts serving as the target approximately twice. The critical question was which matrix's predictions would align more closely with human similarity judgments.

Participants 31 participants were recruited from a University subject pool. Participants completed the task online for course credit. Each participant provided informed consent in compliance with the Institutional IRB.

Stimuli and Procedure. The stimuli were the set of 1,424 triplets described above. Data were collected online via jsPsych (De Leeuw, 2015). On each trial, a randomly selected triplet was displayed, with participants indicating which of two options was more similar to the target concept using a mouse click. All triplets were judged by each participant[2].

$$
\begin{aligned}
P_{human}(i,j,k) &= \frac{d_{ik}^{human}}{d_{ik}^{human} + d_{jk}^{human}}, \\
P_{llm}(i,j,k) &= \frac{d_{ik}^{llm}}{d_{ik}^{llm} + d_{jk}^{llm}}, \\
\text{Discrepancy}(i,j,k) &= -1 \times (P_{human}(i,j,k) - 0.5) \\
&\quad \times (P_{llm}(i,j,k) - 0.5),
\end{aligned}
\tag{1}
$$

Results. Human similarity judgments agreed with predictions of the AI-enhanced norms for 86.20% of triplets, a result unlikely to arise by chance $p < 0.001$, binomial test). Human judgments agreed with predictions of the FastText embeddings on 60.40% of trials: reliably better than chance ($p < 0.001$, binomial test), but significantly worse than the AI-enhanced embeddings (paired t-test, $t(1423)$ =18.37, $p < 0.001$). Thus the richer structure evident in the AI-enhanced feature norms appear to better express human-discerned semantic similarity structure than to norms derived from humans alone or from word-embeddings.

5 Discussion

We present both a new approach for generating AI-enhanced semantic norms along with an accompanying dataset. We first conducted controlled experiments evaluating LLM feature verification performance against a reliable subset of human norm judgments in order to assess and selected the most human-aligned model. We then further enhanced the performance of an open-sourced model by selectively incorporating responses from a more powerful frontier model, GPT-4o, improving norm quality without full reliance on proprietary model outputs. Using our best performing model combination, we generated a large-scale norm dataset spanning over 750 concepts and over 8,000 features. Inspection of the organization of concepts based on these generated features showed that concepts showed a greater degree of feature overlap relative to the raw human-generated matrix. This overlap of features did not come at the cost of category selectivity, with concepts being reasonably organized into meaningful clusters. We further assessed the quality of our norms dataset by conducting a triadic judgment task, the results of which showed that semantic vectors derived from the AI-enhanced matrix more accurately predicted human similarity judgments than those based on human norms alone or word embedding models.

Taken together, our work addresses longstanding limitations in semantic norm generation by creating a dataset that is (1) diverse in the concepts and features represented and (2) verified for feature-concept validity. The feature density of the AI-enhanced norms reveals semantic similarity structure richer than previous norm datasets, unlocking the potential to better understand the neural basis of semantic memory (Clarke & Tyler, 2014; Rogers & McClelland, 2004; Cox et al., 2024; Fernandino et al., 2022) and to guide the development of future computational neurocognitive models (Dilkina et al., 2008; Riordan & Jones, 2011; Saxe et al., 2019; Giallanza et al., 2024; Suresh et al., 2024). Lastly, the present work highlights the promise of integrating large-language models into workflows for cognitive science research in a controlled and verifiable manner and provides a replicable framework for future endeavors in this domain (Suresh et al., 2023a; Mukherjee et al., 2023; Trott, 2024; Dillion et al., 2023; Mukherjee et al., 2024).

[2]there was data loss of a few trials for some participants due to technical issues.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Mark H Ashcraft. Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, 6:227–232, 1978.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Erin M Buchanan, Kathrene D Valentine, and Nicholas P Maxwell. English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, 51:1849–1863, 2019.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *arXiv preprint arXiv:2411.00238*, 2024.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*, 2023a.

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Evaluating llm agent group dynamics against human group dynamics: A case study on wisdom of partisan crowds. *arXiv preprint arXiv:2311.09665*, 2023b.

Alex Clarke and Lorraine K Tyler. Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, 34(14):4766–4775, 2014.

Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

Christopher R Cox, Timothy T Rogers, Akihiro Shimotake, Takayuki Kikuchi, Takeharu Kunieda, Susumu Miyamoto, Ryosuke Takahashi, Riki Matsumoto, Akio Ikeda, and Matthew A Lambon Ralph. Representational similarity learning reveals a graded multidimensional semantic space in the human anterior temporal cortex. *Imaging Neuroscience*, 2:1–22, 2024.

George S Cree and Ken McRae. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of experimental psychology: general*, 132(2):163, 2003.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Simon De Deyne, Steven Verheyen, Eef Ameel, Wolf Vanpaemel, Matthew J Dry, Wouter Voorspoels, and Gert Storms. Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40:1030–1048, 2008.

Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47:1–12, 2015.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46:1119–1127, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Katia Dilkina, James L McClelland, and David C Plaut. A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2):136–164, 2008.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Martha J Farah and James L McClelland. A computational model of semantic memory impairment: Modality specificity and emergent category specificity (journal of experimental psychology: General, 120 (4), 339–357). *Exploring Cognition: Damaged Brains and Neural Networks*, pp. 79–110, 2013.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, 119(6):e2108091119, 2022.

Peter Garrard, Matthew A Lambon Ralph, Peter C Watson, Jane Powis, Karalyn Patterson, and John R Hodges. Longitudinal profiles of semantic impairment for living and nonliving concepts in dementia of alzheimer's type. *Journal of Cognitive Neuroscience*, 13(7):892–909, 2001.

Tyler Giallanza and Declan Iain Campbell. Context-sensitive semantic reasoning in large language models. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Tyler Giallanza, Declan Campbell, Jonathan D Cohen, and Timothy T Rogers. An integrated model of semantics and control. *Psychological Review*, 2024.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*, 2023.

Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1, 2023.

Hannes Hansen and Martin N Hebart. Semantic features of object concepts generated with gpt-3. *arXiv preprint arXiv:2202.03753*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Abhilasha A Kumar. Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28(1):40–80, 2021.

Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. Words are all you need? capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105*, 2022.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*, 2023.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547, 2005.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. *arXiv preprint arXiv:2210.01963*, 2022.

Kushin Mukherjee, Siddharth Suresh, and Timothy T Rogers. Human-machine cooperation for semantic feature listing. *arXiv preprint arXiv:2304.05012*, 2023.

Kushin Mukherjee, Timothy T Rogers, and Karen B Schloss. Large language models estimate fine-grained human color-concept associations. *arXiv preprint arXiv:2406.17781*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Ankit Patel. NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models. `https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/`, 2024. Accessed: 2025-02-05.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Brian Riordan and Michael N Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345, 2011.

Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

Wim Ruts, Simon De Deyne, Eef Ameel, Wolf Vanpaemel, Timothy Verbeemen, and Gert Storms. Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36(3):506–515, 2004.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*, 2024.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Siddharth Suresh, Kushin Mukherjee, and Timothy T Rogers. Semantic feature verification in flan-t5. *arXiv preprint arXiv:2304.05591*, 2023a.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 722–738, 2023b.

Siddharth Suresh, Wei-Chun Huang, Kushin Mukherjee, and Timothy T Rogers. Categories vs semantic features: What shape the similarities people discern in photographs of objects? In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Sean Trott. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, pp. 1–19, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

A Appendix

A.1 Detailed Explanation of the Triplet Discrepancy Metric

For any triplet of concepts with target i and two candidate options j and k, we quantify the similarity relationships in two representational spaces (e.g., human judgments and a language model) via relative distance measures. Specifically, for each space we compute a probability that reflects the extent to which the distance between the target and one option outweighs that of the alternative. Formally, we define:

$$P_{\text{human}}(i,j,k) = \frac{d_{ik}^{\text{human}}}{d_{ik}^{\text{human}} + d_{jk}^{\text{human}}}, \quad P_{\text{llm}}(i,j,k) = \frac{d_{ik}^{\text{llm}}}{d_{ik}^{\text{llm}} + d_{jk}^{\text{llm}}}. \tag{2}$$

Here, d_{ik}^{human} (or d_{ik}^{llm}) denotes the distance between concepts i and k in the human (or language model) similarity space. These probabilities lie in the interval $[0, 1]$, with a value of 0.5 indicating indifference (i.e., no preference for one option over the other).

To assess the confidence and direction of the preference, we center these probabilities by subtracting 0.5. A positive deviation, $P - 0.5$, implies a preference for one option, while a negative deviation implies a preference for the alternative.

Our primary quantity of interest is the *discrepancy* between the two spaces, defined as:

$$\text{Discrepancy}(i,j,k) = -\Big(P_{\text{human}}(i,j,k) - 0.5\Big) \cdot \Big(P_{\text{llm}}(i,j,k) - 0.5\Big). \tag{3}$$

This formulation has two key properties:

1. **Sign of the Discrepancy:**
   - When both $P_{\text{human}}(i,j,k)$ and $P_{\text{llm}}(i,j,k)$ deviate from 0.5 in the same direction (i.e., both systems agree on which option is more similar to the target), the product
   $$\Big(P_{\text{human}}(i,j,k) - 0.5\Big)\Big(P_{\text{llm}}(i,j,k) - 0.5\Big)$$
   is positive, and thus the discrepancy becomes negative after multiplication by -1.
   - Conversely, if one probability is above 0.5 and the other below 0.5 (i.e., the two systems disagree), the product is negative, and the discrepancy becomes positive after applying the negative sign.
2. **Magnitude of the Discrepancy:**
   - The magnitude of each term $|P - 0.5|$ reflects the degree of confidence in the respective judgment. Thus, triplets where both the human and LLM judgments are made with high confidence (i.e., probabilities far from 0.5) will yield a larger absolute discrepancy value.
   - In contrast, when either system exhibits little bias (i.e., P is near 0.5), the resulting discrepancy will be small, even if the signs are opposed.

Triplets with large positive discrepancy values therefore represent cases where the human judgments and the LLM predictions are both confident yet in direct conflict. Selecting these triplets allows us to focus our analysis on the stimuli that reveal the most substantial disagreements between the two representational spaces.

This metric thus serves as an effective tool for identifying and prioritizing triplets that are most informative for understanding the divergence between human and model similarity assessments.