Behavioral estimates of conceptual structure are robust across tasks in humans but not large language models

Siddharth Suresh[1] **(siddharth.suresh@wisc.edu)**
Lisa Padua[2] (lpadua@students.asurams.edu)
Kushin Mukherjee[1] (kmukherjee2@wisc.edu)
Timothy T. Rogers[1] (ttrogers@wisc.edu)

Abstract

Neural network models of language have long been used as a tool for developing hypotheses about conceptual representation in the mind and brain. For many years, such use involved extracting vector-space representations of words and using distances among these to predict or understand human behavior in various semantic tasks. In contemporary language AIs, however, it is possible to interrogate the latent structure of conceptual representations using methods nearly identical to those commonly used with human participants. The current work uses two common techniques borrowed from cognitive psychology to estimate and compare lexical-semantic structure in both humans and a well-known AI, the DaVinci variant of GPT-3. In humans, we show that conceptual structure is robust to differences in culture, language, and method of estimation. Structures estimated from AI behavior, while individually fairly consistent with those estimated from human behavior, depend much more upon the particular task used to generate behavior responses–responses generated by the very same model in the two tasks yield estimates of conceptual structure that cohere less with one another than do human structure estimates. The results suggest one important way that knowledge inhering in contemporary AIs can differ from human cognition.

Keywords: Artificial Intelligence, Semantic memory, Natural Language Processing, Knowledge representation, Neural Networks

Introduction

Since Elman's pioneering work(J. L. Elman, 1990) showcasing the ability of neural networks to capture the rich statistics of human natural language through backpropagation, these models have provided a useful tool, and sometimes a gadfly, for developing hypotheses about the cognitive and neural mechanisms that support language. When trained on a task that seems almost absurdly simplistic–continuous, sequential prediction of upcoming words in sentences–early models exhibited properties that upended received wisdom about what language is and how it works. They acquired internal representations that blended syntactic and semantic information, rather than keeping these separate as classic psycholinguistics required. They handled grammatical dependencies, not by constructing syntactic structure trees, but by learning and exploiting temporal patterns in language. Perhaps most surprisingly, they illustrated that statistical structure latent in lexical contexts could go a long way toward explaining how we acquire knowledge of semantic similarity relations among words. Because words with similar meanings tend to be encountered in similar linguistic contexts(Firth, 1957), models that exploit contextual similarity when representing words come to express semantic relations between them.

Though early work was limited in the nature and complexity of the "language" used to train the models(J. Elman, 1991; McClelland, Rumelhart, & the PDP Research Group, 1990), these ideas spurred a variety of computational approaches that could be applied to large corpora of written text. Approaches such as latent semantic analysis(Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and skipgram models(Mikolov, Chen, Corrado, & Dean, 2013), for instance, learn vector-space representations of words from overlap in their linguistic contexts, which turn out to capture a variety of semantic relationships amongst words, including some that are highly abstract(Grand, Blank, Pereira, & Fedorenko, 2022; J. L. Elman, 2004; Lupyan & Lewis, 2019).

In all of this work, lexical-semantic representations are cast as static points in a high-dimensional vector space, either computed directly from estimates of word co-occurrence in large text corpora(Deerwester et al., 1990; Burgess, 1998), or instantiated as the learned activation patterns arising in a neural network model trained on such corpora. To evaluate whether a given approach expresses semantic structure similar to that discerned by human participants, the experimenter typically compares the similarities between word vectors learned by a model to decisions or behaviors exhibited by participants in semantic tasks. For instance, LSA models were tested on synonym-judgment tasks drawn from a common standardized test of English language comprehension by comparing the cosine distance between the vectors corresponding to a target word and each of several option words, and having the model "choose" the option with the smallest distance (Landauer, Foltz, & Laham, 1998). The model was deemed successful because the choice computed in this way often aligned with the choices of native English speakers. Such a procedure was not just a useful way for assessing whether model representations are human-like—it was just about the only way to do so for this class of models. That is, we were limited in the ways we could interact with the model — one couldn't just treat the model like a human participant by giving it a set of instructions and a series of stimuli and

[1] Department of Psychology, University of Wisconsin-Madison
[2] Department of Psychology, Albany State University

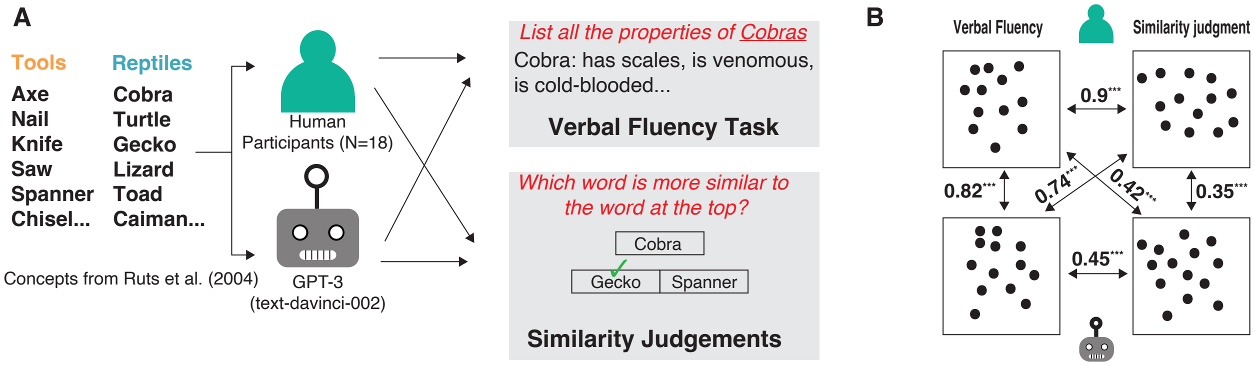


Figure 1: (A) Procedure for querying both human and GPT3 conceptual structure. Both agents performed the same verbal fluency and triplet similarity judgement task. The data generated from these tasks were used to fit low-dimensional feature vectors. (B) The squared procrustes correlation between embedding spaces obtained from the Verbal Fluency Task on humans (Human feat lists), Similarity Judgements on humans (Human triplets), Verbal Fluency on GPT-3 (GPT feat. lists(verified)), and Similarity Judgements on GPT-3 (GPT triplets).

just recording its responses.

In the era of large language models such as Open AI's GPT3 (Brown et al., 2020), Meta's OPT 175(Zhang et al., 2022), Google's FLAN (Wei et al., 2021), and others(Chowdhery et al., 2022; Hoffmann et al., 2022; Du et al., 2022), this has changed. Such models are many orders of magnitude larger than classical connectionist approaches, employ a range of architectural and training innovations, and are optimized on truly fast quantities of data—but nevertheless they operate on principles not dissimilar to those that Elman and others pioneered. That is, they exploit patterns of word co-occurrence in natural language to learn distributed, context-sensitive representations of linguistic meaning at multiple levels, and from these representations they generate probabilistic predictions about likely upcoming words. Given an initial linguistic prompt, such models generate plausible and grammatically well-formed responses, created by iteratively predicting what words are likely to come next and sampling from this distribution. The results are not only eerily human-like; recent iterations like Chat GPT (Ouyang et al., 2022) can write essays sufficient to earn a B in a typical college class, produce working Python code from a general description of the function, generate coherent explanations for a variety of phenomena, and answer factual questions with remarkable accuracy. In short, such models appear to show several hallmarks of conceptual abilities that until recently were uniquely human.

These innovations allow cognitive scientists, for the first time, to measure and evaluate conceptual structure in a non-human system using precisely the same natural-language-based methods that we use to study human participants. Large language models can receive written instructions followed by a series of stimuli and generate interpretable, natural-language responses for each. The responses generated can be recorded and analyzed in precisely the same manner as responses generated by humans, and the results of such analyses can then be compared within and between human and AI systems, as a means of understanding whether and how these intelligences differ.

The current paper uses this approach to understand similarities and differences in the way that lexical semantic representations are structured in human and AI minds, focusing on one remarkable aspect of human concepts–specifically, their *robustness*. As Rosch showed many years ago(E. Rosch, 1975; E. H. Rosch, 1973), the same conceptual relations underlie behavior in a variety of tasks, from naming and categorization to feature-listing to similarity judgments to sorting. Similar conceptual relations can be observed across distinct languages and cultures(Thompson, Roberts, & Lupyan, 2020). Robustness is important because it allows for shared understanding and communication across cultures, over time, and through generations: Homer still speaks to us despite the astonishing differences between his world and ours, because many of the concepts that organized his world cohere with those that organize ours. Our goal was to assess whether conceptual structure in a contemporary AI (a Large Language Model) is also coherent when evaluated using methods comparable to those employed with human participants, or whether human and AI "minds" differ in this important regard.

To answer this question, we first measured the robustness of conceptual structure in human agents by comparing estimates of such structure for a fixed set of concepts using two different behavioral methods in two distinct groups differing in both culture and language. We then conducted the same behavioral experiments on a large-language AI (the DaVinci variant of GPT3), and evaluated (a) the degree to which estimated conceptual relations in the AI accord with those observed in humans, and (b) whether humans and AI differ in the apparent robustness of such structure. We further compared the structures estimated from the AI's overt patterns of behavior to those encoded in its internal representations, and also to semantic vectors extracted from two other common models in machine learning. In addition to simply demonstrating how methods from cognitive psychology can be used to better understand machine intelligence, the results point to an important difference between current state of the art AI and human conceptual representations.

Study 1: How coherent is human conceptual structure?

The goal of study 1 was to evaluate how robust estimates of conceptual structure appear when generated from human behavior in two very different populations and tasks, using methods that can be replicated with large language models. To this end, we focused on a subset of 30 concepts taken from a large feature-norming study conducted at KU Leuven (De Deyne et al., 2008). We estimated semantic similarity relations amongst these items using two distinct methods: first via vector distance in the feature space produced, as captured by the norms themselves, and second using a triadic-comparison task to estimate low-dimensional embeddings for the words that express their semantic relatedness.

The resulting datasets differ from each other in (1) the task used (feature generation vs semantic similarity judgments), (2) the language of instruction and production (Dutch vs English), and (3) the population from which the participants were recruited (Belgian students in early 2000's vs American MTurk workers in 2022). The central question was how similar the resulting estimated structures are to one another, a metric we call *structure coherence*. If estimated conceptual similarities vary substantially with language, culture, or estimation method, the structural coherence between groups will be relatively low; if such estimates are robust to these factors, it will be high. The comparison then provides a baseline against which to compare structural coherence in the AI.

Methods

Feature listing study Data were taken from the Leuven large feature-listing norms(De Deyne et al., 2008). In an initial *generation* phase, this study asked 1003 participants to list 10 semantic features for 6-10 different stimulus words which were were one of 295 (129 animals and 166 artifacts) concrete object concepts. The set of features produced across all items were tabulated into a 2600d feature vector. In a second *verification* phase, four independent raters considered each concept-feature pair and evaluated whether the feature was true of the concept. The final dataset thus contained a C (concept) by F (feature) matrix whose entries indicate how many of the four raters judged concept C to have feature F. Note that this endeavour required the raters to judge hundreds of thousands of concept-property pairs!

From the full set of items, we selected 15 tools and 15 reptiles for use in this study. The reptiles were: turtle, alligator, lizard, tortoise, cobra, snake, blindworm, gecko, boa python, toad, crocodile, chameleon, caiman, salamander, and dinosaur. The tools were: hammer, screwdriver, grinding disc, vacuum cleaner, spanner, lawn mower, axe, saw, knife, nail, chisel, shovel, anvil, oilcan, paint brush. We chose these categories because they express both broad, superordinate distinctions (living/nonliving) as well as finer-grained internal structure (e.g. snakes vs lizards vs crocodilians).

The raw feature vectors were binarized by converting all non-zero entries to 1, with the rationale that a given feature is potentially true of a concept if at least one rater judged it to be so. Following Rosch (E. H. Rosch, 1973), McRae (McRae, Cree, Seidenberg, & McNorgan, 2005) among others, we then estimated the conceptual similarity relations amongst all pairs of items by taking the cosine distance between their binarized feature vectors, and reduced the space to three dimensions via classical multidimensional scaling (Kruskal & Wish, 1978). The resulting embedding expresses conceptual similarity amongst 30 concrete objects, as estimated via semantic feature listing and verification, in a study conducted in Dutch on a large group of students living in Belgium in the early 2010s.

Triadic comparison study. As a second estimate of conceptual structure amongst the same 30 items, we conducted a triadic comparison or *triplet judgment* task in which participants must decide which of two option words is more similar in meaning to a third reference word. From many such judgments, ordinal embedding techniques(Jamieson, Jain, Fernandez, Glattard, & Nowak, 2015) can be used to situate words within a low-dimensional space in which Euclidean distances between two words capture the probability that they will be selected as "most similar" relative to some arbitrary third word. Like feature-listing, triplet judgment studies can be conducted completely verbally, and so can be simulated using large language models.

Participants were 18 Amazon Mechanical Turk workers recruited using CloudResearch. Each participant provided informed consent in compliance with our Institutional IRB and was compensated for their time.

Stimuli were English translations of the 30 item names listed above, half reptiles and half tools.

Procedure. On each trial participants viewed a target word displayed above two option words, and were instructed to choose via button press which of the two option words was most similar to the target in its meaning. Each participant completed 200 trials, with the triplet on each trial sampled randomly with uniform probability from the space of all possible triplets. The study yielded a total of 3600 judgments, an order of magnitude larger than the minimal needed to estimate an accurate 3D embedding from random sampling according to estimates of sample complexity in this task(Jamieson et al., 2015). Ninety percent of the judgments were used to find a 3D embedding in which pairwise Euclidean distances amongst words minimize the crowd-kernel triplet loss on the training set(Tamuz, Liu, Belongie, Shamir, & Kalai, 2011). The resulting embedding was then tested by assessing its accuracy in predicting human judgments on the held-out ten percent of data. The final embeddings pre-

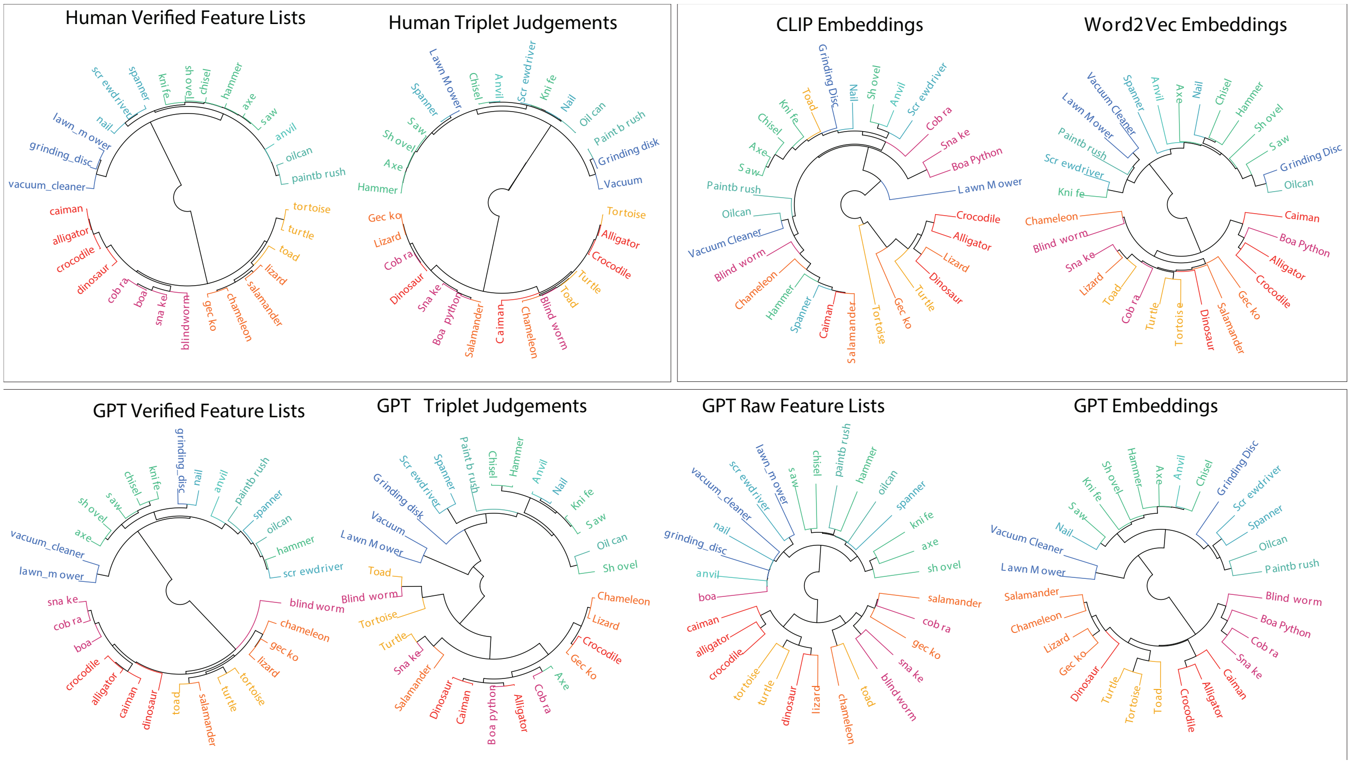


Figure 2: Hierarchical cluster plots showing how concepts are organized within each semantic feature space. (Top left) Spaces generated from human data, (top right) spaces generated from neural network embeddings, and (bottom) spaces generated from both GPT3 behavioral results and GPT3 embeddings.

dicted human decisions on held-out triplets with 75% accuracy, which matched the mean level of inter-subject agreement on this task.

Results

Figure 2 top left shows hierarchical cluster plots of the semantic embeddings from feature lists (left) versus the triadic comparison task (right). Both embeddings strongly differentiate the living from nonliving items, and show comparatively little differentiation of subtypes within each category (though such subtypes are clearly apparent amongst the feature-listing embeddings). To estimate how structurally coherent the two different embedding spaces are, we computed the square of the Procrustes correlation (Gower, 1975) between the two 3D embeddings, a metric analogous to r^2 that indicates how much of the variation in pairwise distances from one matrix is captured by distances in the other. This metric was 0.90, very reliably better than chance ($p < 0.001$) and indicating that distances in one space capture 90% of the variation in the other. Thus despite differences in language, task, and cultures, the two estimates of conceptual structure were well-aligned, suggesting that human conceptual representations of concrete objects are remarkably robust. We next consider whether the same is true of large language models.

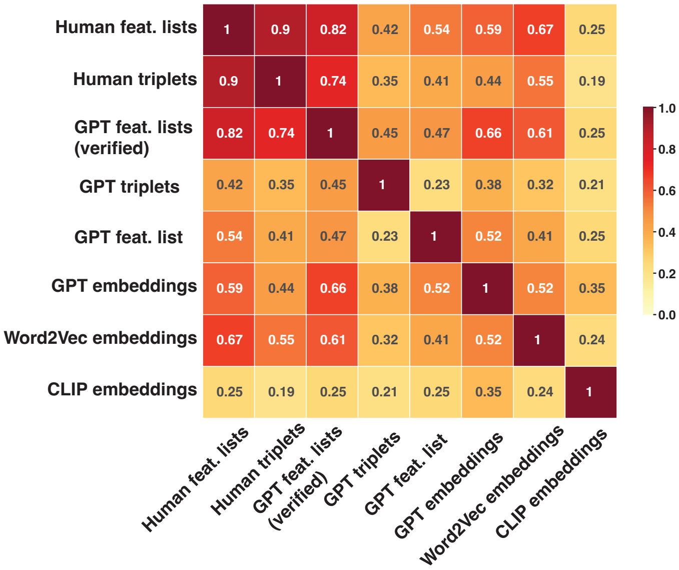


Figure 3: Correlation structure between different semantic feature vectors. Feature vectors were generated from human and machine behavioral tasks as well as extracted from standard NLP and multimodal models. All the values are significant with $p < 0.001$

Study 2: Evaluating structural coherence of concepts within and between AI and human.

Study 2 aimed to estimate conceptual similarity relations for the same 30 items from the behavior of an AI (the DaVinci variant of GPT-3) asked to perform feature-listing and triadic-judgment tasks analogous to those completed by human participants. We first developed and conducted both tasks on GPT-3 using their API, and computed semantic embeddings from the text generated by the model using the same techniques employed with human data. We then considered (a) how well these estimates aligned with structures estimated from human behaviors within each task, and (b) the structural coherence between the two embeddings estimated via different methods from AI behavior.

Methods

Feature listing in GPT-3 To simulate the feature-generation phase of the Leuven study, We queried GPT-3 with the prompt "List the features of a [concept]" and recorded the responses (see Figure 4 left). The model was queried with a temperature of 0.7, meaning that responses were somewhat stochastic so that the model produced different responses from repetitions of the same query. For each concept We repeated the process five times and tabulated all responses across these runs for each item. The responses were transcribed into features by breaking down phrases or sentence into constituent predicates; for instance, a response such as "a tiger is covered in stripes" was transcribed as "has stripes." Where phrases included modifiers, these were transcribed separately; for instance, a phrase like "has a long neck" was transcribed as two features, "has neck" and "has long neck." Finally, alternate wordings for the same property were treated as denoting a single feature; for instance, "its claws are sharp" and "has razor-sharp claws" would be coded as the two features "has claws" and "has sharp claws." We did not, however, collapse synonyms or otherwise reduce the feature set. This exercise generated a total of 580 unique features from the 30 items.

To simulate the feature verification phase of the Leuven study, we then asked GPT to decide, for each concept C and feature F, whether the concept possessed the feature (Figure 4 middle). For instance, to assess whether the model "thinks" that alligators are ectothermic, we probed it with the prompt "In one word, Yes/No: Are alligators ectothermic?" (temperature 0). Note that this procedure requires the AI to answer probes for every possible concept/feature pair–for instance, does an alligator have wheels? Does a car have a heart? etc. These responses were used to flesh out the original feature-listing matrix: every cell where the AI affirmed that concept C had feature F was filled with a 1, and cells where the AI responded "no" were filled with zeros. We refer to the resulting matrix as the *verified feature matrix*. Before the feature verification process, the concept by feature matrix was exceedingly sparse, containing 786 1's (associations) and 16614 0's (no associations). After the verification process, the concept by feature matrix contained 7845 1's and 9555 0's. Finally, we computed pairwise cosine distances between all items based on the verified feature vectors, and used classical multidimensional scaling to reduce these to three-dimensional embeddings, exactly comparable to the human study.

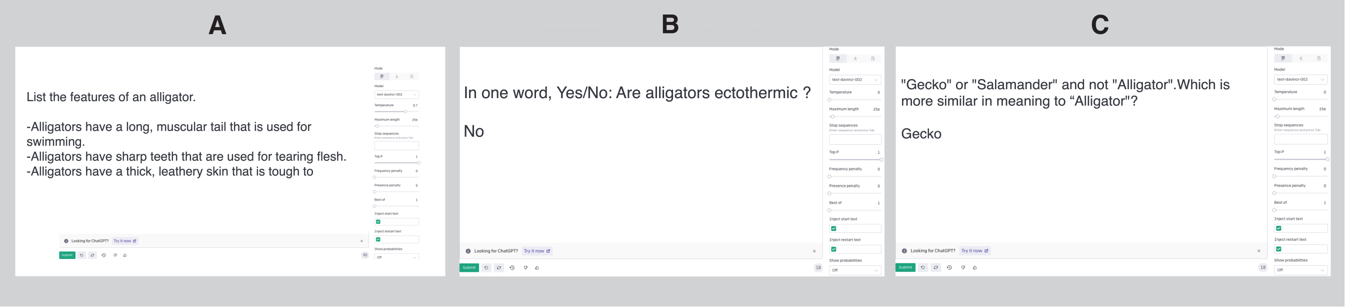
Triplet judgment in GPT-3. To simulate triplet judgment, we used the prompt shown in Figure 4 (right) for each triplet, using the exact same set of triplets employed across all participants in the human study. We recorded the AI's response for each and from these data fit a 3D embedding using the same algorithm and settings as the human data. The resulting embedding predicted GPT-3 judgements for the held-out triplets at a 78 % accuracy, comparable to that observed in the human participants.

Results

Hierarchical cluster plots for embeddings generated from the AI's feature lists and triplet judgments are shown in the bottom left panels of Figure 2, immediately below the corresponding plots from human data. Both approaches reliably separate living and nonliving things. The verified feature lists additionally yield within-domain structure similar to that observed in human lists, with all items relatively similar to one another, and with some subcategory structure apparent (e.g. turtle/tortoise, snakes, crocodilia). within-domain structure estimated from triplet judgments, in contrast, looks very different.

These qualitative observations are borne out by the squared Procrustes correlations between different embedding spaces, shown in Figure 3. Similarities expressed in the space estimated from verified feature lists capture 82% of the variance in distances estimated from human feature lists, and 74% of the variance in those estimated from human triplet judgments. Similarities estimated from AI triplet judgments, in contrast, account for less than half the variance in embeddings generated from human judgments. More interestingly, they account for less than half the variance in the embeddings generated from the AI verified feature lists. Unlike the human embeddings, conceptual structures estimated from different behaviors in the very same model do not cohere very well with each other.

Figure 3 also shows the squared Procrustes correlations for semantic embeddings generated via several other approaches including (a) the raw (unverified) feature lists produced by GPT-3, (b) the word embedding vectors extracted from GPT-3's internal hidden unit activation patterns, (c) word embeddings from the popular word2vec approach, and (d) embeddings extracted from a CLIP model trained to connect images with their natural language descriptions. None of these approaches accord with human-based embeddings as well as do the embeddings estimated from the AI verified-feature lists, nor are the various structures particularly coherent with one another. No pair of AI-estimated embeddings shows the degree of coherence observed between the estimates derived from human judgments.



A

List the features of an alligator.

-Alligators have a long, muscular tail that is used for swimming.
-Alligators have sharp teeth that are used for tearing flesh.
-Alligators have a thick, leathery skin that is tough to

Model text-davinci-002
Temperature 0.7
Maximum length 256
Stop sequences
Top P 1
Frequency penalty 0
Presence penalty 0
Best of 1
Inject start text ☑
Inject restart text ☑
Show probabilities Off

B

In one word, Yes/No: Are alligators ectothermic ?

No

Model text-davinci-002
Temperature 0
Maximum length 256
Stop sequences
Top P 1
Frequency penalty 0
Presence penalty 0
Best of 1
Inject start text ☑
Inject restart text ☑
Show probabilities Off

C

"Gecko" or "Salamander" and not "Alligator".Which is more similar in meaning to "Alligator"?

Gecko

Model text-davinci-002
Temperature 0
Maximum length 256
Stop sequences
Top P 1
Frequency penalty 0
Presence penalty 0
Best of 1
Inject start text ☑
Inject restart text ☑
Show probabilities Off

Figure 4: Prompts for querying GPT-3 to perform (A) Feature generation, (B) Feature Verification and (C) Similarity judgement.

Discussion

In this study, we compared the conceptual structures of humans and GPT-3 using two cognitive tasks: a semantic feature-listing task and a triplet similarity judgement task. Our results showed that the conceptual representations generated from human judgments, despite being estimated from quite different tasks, in different languages, across different cultures, were remarkably coherent: similarities captured in one space accounted for 90% of the variance in the other. This suggests that the conceptual structures underlying human semantic cognition are remarkably robust to differences in language, cultural background, and the nature of the task at hand.

In contrast, embeddings obtained from analogous behaviors in GPT-3 differed depending upon on the task. While embeddings estimated from verified feature lists aligned moderately well with those estimated from human feature norms, those estimated from triplet judgments or from the raw (unverified) feature lists did not, nor did the two embedding spaces from the AI cohere well with each other. Embedding spaces extracted directly from model hidden representations or from other common neural network techniques did not fare better: in most comparisons, distances captured by one model-derived embedding space accounted for, at best, half the variance in any other. The sole exception was the space estimated from AI verified feature vectors, which cohered modestly well with embeddings taken directly from the GPT-3 hidden layer (66% of variance) and with word2vec embeddings (61%).

Together these results suggest an important difference between human cognition and current AI models. Neurocomputational models of human semantic memory suggest that behavior across many different tasks is undergirded by a common conceptual ”core” that is relatively insulated from variations arising from different contexts or tasks(Rogers, McClelland, et al., 2004; Jackson, Rogers, & Lambon Ralph, 2021). In contrast, representations of word meanings in large language models depend essentially upon the broader linguistic context. Indeed, in transformer architectures like GPT3, each word vector is computed as a weighted average of vectors from surrounding text, so it is unclear whether any word possesses meaning outside or independent of context. Because this is so, the latent structures organizing its overt behaviors may vary considerably depending upon the particular way the model's behavior is probed. That is, the AI may not have a coherent conceptual ”core” driving its behaviors, and for this reason, may organize its internal representations quite differently with changes to the task instruction or prompt. Context-sensitivity of this kind is precisely what grants such models their notable ability to simulate natural-seeming language, but this same capacity may render them ill-suited for understanding human conceptual representation.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the hal model. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 188–198.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... others (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, *40*, 1030–1048.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022). Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 320–335).

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2-3), 195–225.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in cognitive sciences*, *8*(7), 301–306.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*, 33–51.

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.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., … others (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature human behaviour*, *5*(6), 774–786.

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. D. (2015). Next: A system for real-world development, evaluation, and application of active learning. In *Nips* (pp. 2656–2664).

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (Vol. 11). Sage.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

McClelland, J., Rumelhart, D., & the PDP Research Group. (1990). The development of distributed representations for words. *Explorations in parallel distributed processing: a handbook of models, programs, and exercises*, 77–109.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *International conference on machine learning* (pp. 1188–1196).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … others (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Rogers, T. T., McClelland, J. L., et al. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

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

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language* (pp. 111–144). Elsevier.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029–1038.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., … others (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.