Evaluating human-like similarity biases at every scale in Large Language Models: Evidence from remote and basic-level triads.

Simon De Deyne (simon.dedeyne@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Australia

Abstract

In the remote triad task, participants judge the relatedness between randomly chosen words in a three-alternative choice triadic judgement task. While most word pairs in these triads are weakly related, humans agree on which to choose. This is theoretically interesting as it contradicts previous claims that suggest that the notion of similarity is unconstrained in principle (e.g., Goodman, 1972). Here, we present new evidence from GPT-4, showing that context-aware LLMs provide excellent predictions of this task. Moreover, the strength of this effect was even larger than that found for basic-level comparisons, which involve highly similar items. Together, this implies that the similarity of human representations is highly structured at every scale, even in tasks with limited context. Follow-up analysis provides insights into how LLMs are successful in this task. Further implications of the ability to compare words at every scale are discussed.

Keywords: word meaning; concepts, remote triads, basic-level comparisons, GPT-4

Introduction

Similarity is a key notion in cognitive science, explaining a range of phenomena related to concepts, word meaning, categorization and theories of memory. However, similarity has been criticised as a theoretical construct for being too flexible and underspecified to ground cognition. Any pair of objects can be similar in infinite ways unless one specifies in what respects two things are similar (Goodman, 1972). The problem of comparing things without these respects is nicely illustrated in idioms about sensible comparisons that seem to be universally shared among many languages (*apples and pears* in Dutch; *chalk and cheese* in English; *shoes and wheels* in Lithuanian; *gingerbread and windmills* in Polish).

Perhaps surprisingly, empirical work demonstrated that similarity is sufficiently constrained, with humans exhibiting a large degree of agreement in judging the similarity in relatively context-free tasks, including judging items from a category (e.g. vegetables) and even weakly related concepts randomly sampled from a list of words (De Deyne, Navarro, Perfors, & Storms, 2012; De Deyne, Navarro, Collell, & Perfors, 2021; Richie & Bhatia, 2021). In theory, language-based models trained on enormous corpora should be able to capture the same weak relations as they can generalise weak contingencies between words by predicting how words can co-occur. Several standard benchmarks where humans are asked to judge the similarity of word pairs suggest this is the case (Baroni, Dinu, & Kruszewski, 2014; Mandera, Keuleers, & Brysbaert, 2017). However, more challenging datasets that cover the range of similarity between (lexicalised) concepts systematically show that these models struggle to capture the relations between weakly related (remote) concepts and perceptually rich basic-level comparisons for abstract and concrete concepts that are close(De Deyne et al., 2021). Together, these findings suggest that commonly used similarity benchmarks are not always sufficiently discriminating because the range of similarity is inflated by including both highly similar and highly dissimilar items, thus boosting prediction metrics. In contrast, the results for more challenging datasets suggest that models explain only a fraction of the variance (Richie & Bhatia, 2021; De Deyne, Perfors, & Navarro, 2016).

This study investigates whether Transformer-based Large Language Models capture similarity at every scale, based on new evidence demonstrating that recent transformer-based models like GPT-4 (see Brown et al., 2020) outperform word embedding models (Trott, 2023). We do so by re-analysing two landmark studies that draw similarity judgments at extreme ends of the scale. The first data set is from two studies that used a remote triad task. In this task, word triples are sampled randomly, and participants are asked which pair is strongest related. Due to the random sampling, most words only have weak similarity relationships. Here, we ask whether weak similarity is sufficiently encoded in language, thus providing a test of a more moderate of Goodman's argument where similarity might be less constrained if items are less related. The second set compares basic-level members of common concrete and abstract categories such as fruits, vehicles, emotions or sciences. This is an example of a challenging task at the other (closer) end of the scale since humans might rely on extralinguistic information, such as perceptual and affective information, when judging similarity (Richie & Bhatia, 2021; De Deyne et al., 2021). Beyond covering the extreme ends of the scale, these tasks were selected for two additional reasons. First, they have previously provided insight into the limitations of distributional semantic models derived from text corpora by providing a challenging test. Second, triadic comparisons used in these tasks have some advantages over other methods like pairwise ratings as they don't require the use of absolute scale but offer a relative comparison between pairs anchored by a third option, which might be a discriminating feature between context-free

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

and context-aware representations in word embeddings and Transformer architectures. Below, we briefly review previous findings for both tasks.

Remote triads In the remote triad tasks, participants are asked to identify the most related pair in a triad with randomly selected options. Because the words in each triad were randomly sampled, most of them are only weakly related (e.g. *hyena – radish – somersault*). Previous work showed a surprising degree of agreement among participants, and this finding was replicated in both Dutch (De Deyne et al., 2012) and English (De Deyne et al., 2021). In both studies, a semantic network model from word association data successfully predicted human preferences, even though the words in the triads were not directly associated, which suggests that method overlap is unlikely to explain these findings (see also De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). Key to this prediction was a spreading activation mechanism based on random walks to address the sparsity of the input, as each word only has a relatively small set of distinct associates. In contrast, predicted weak contingencies between words in word embedding models did not seem to capture these data as well, with word2vec correlating only $r = .52$ compared to $r = .74$ for the semantic network approach derived from a much smaller dataset (De Deyne et al., 2016).

Basic-level triads The basic level is the most informative taxonomic level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). At this level, concrete concepts like *cat-dog* or *cherry-peach* are predominantly grouped into categories based on perceptual overlap. Previous work has shown that language models capture this perceptual information only partially, and the prediction of word embedding models can be significantly improved by adding perceptual information. For example, for the basic-level triads data in De Deyne et al. (2021), adding visual features derived from a ResNet, a CNN trained on ImageNet, improved performance from $r = .64$ to $r = .75$.

In contrast to concrete concepts, abstract concepts lack a perceivable referent and are often assumed to be acquired primarily through language. They are characterized by their links to other concepts rather than intrinsic properties they may have (Borghi et al., 2017), making them highly relational. While there have been many recent theories about the representation of abstract concepts (Borghi et al., 2017), several theories highlight the role of internal affective states to ground abstract concepts (Vigliocco, Meteyard, Andrews, & Kousta, 2009). Like concrete concepts, cues to affective information are often encoded extra-linguistically. In a basic-level triadic comparison task with abstract concepts, De Deyne et al. (2021) found that adding affective information significantly improved the prediction of human data, from $r = .62$ to $r = .74$. Altogether, this suggests that for both concrete and abstract concepts, basic level comparisons are challenging for word embedding models for different reasons, even though the words at this end of the scale share a relatively large number of features.

Current study Based on recent literature, we expect LLMs to improve significantly over word embedding models. This would be consistent with their superior performance on other similarity datasets that are highly discriminative, such as SimLex and SimVerb, which require strict judgments of similarity (Trott, 2023) by asking participants to consider related items such as *dog - bone* or antonyms such as *black-white* as dissimilar. It would also be supported by work showing the ability of models like GPT-4 to capture modal-specific information (Marjieh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2023), which would be crucial in predicting basic-level concrete triads. However, other studies show limitations as well. In a study by Han, Ransom, Perfors, and Kemp (2024), GPT-4 was evaluated using a large-scale pairwise similarity judgment task with basic-level concepts. GPT-4 outperformed previous models, but the performance was far from perfect (ρ between .38 and .60 for *vegetables*, *reptiles*, *insects*, *clothing*, *fruit*, *sports*, *professions*, *birds*, *tools*, *fish*, *music instruments*, *mammals*, *kitchen utensils*), and only three categories with $\rho > .60$ (*mammals*, *vehicles*, *weapons*).

Furthermore, it is unclear whether LLMs would capture the relation between weakly related concepts, and we consider this a strong test to see if these models encode the same biases that help humans constrain this problem.Given that the reliability of human judgments in the studies included here is high, with split-half reliability $> .90$, a considerable amount of systematic variation remains unaccounted for. Of course, even though models like GPT-4 are trained on massive datasets, including perceptual information, it is unclear whether these data capture human experience if we assume that meaning is also derived from extra-linguistic emotional and perceptual data. It is equally unclear if current models require additional assumptions to explain these data, for example, using different similarity comparisons or category-specific weightings as in Richie and Bhatia (2021). In short, beyond contributing quantitative improvements in prediction, the current evaluation aims to give insight into what kind of theories we need. The second part of the paper aims to work towards this goal of enhancing our understanding of how Large Language Models approach these tasks by comparing the role of similarity versus relatedness and pairwise versus triadic judgments. This will provide insight into whether the model is sensitive to specific instructions and comparisons, which is important as researchers have previously argued that a single representation might not be sufficient to capture both (Mandera et al., 2017). Comparing similarity vs relatedness thus offers an insight into the ability to modulate the role of similarity or relatedness in Large Language models.

Remote and Basic-level Triads: Empirical data

Below, we summarise the main characteristics of the datasets used in this work. Full details can be found in the original papers (De Deyne et al., 2016, 2021).

Participants. The participants of the remote triad tasks consisted of 40 fluent English described in De Deyne et al. (2016). The participants for the basic-level triad tasks were taken from De Deyne et al. (2021) and comprised two groups of 40 native English speakers. The first group judged concrete basic category triads, whereas the second group judged abstract basic category triads (De Deyne et al., 2021).

Materials and Measures. The English stimuli comprised 300 nouns grouped into 100 triads. All items in a triad had similar word frequency and concreteness and were not directly associated according to word association norms. The words were otherwise randomly selected from a large set of words from the USF (Nelson, McEvoy, & Schreiber, 2004) and Small World of Words (SWOW) word association norms (De Deyne et al., 2019).

Procedure. Participants were shown the triads on the corner of an equilateral triangle and were instructed to select the most related pair out of three words displayed on the screen, illustrated by a few examples. Importantly, they were told that the goal is to evaluate the meaning of these words, not the similarity between other things like letters or rhyme, think of relatedness broadly and respond to the best of their abilities even if relatedness was very weak (see De Deyne et al., 2012, for full instructions).

Triadic comparisons

We used the most current GPT-4 version (0613) available at the time of writing with a temperature = 0 through the OpenAI API.[1][2] The triadic comparison prompts were consistent with best practices inspired by previous work by Han et al. (2024). We aimed to keep the instructions consistent between participants and prompts, and explain below where we deviated. Just like the humans, we also included examples to calibrate the model. We modified the instructions slightly to provide feedback on the examples and step-by-step instructions on performing the task and by asking the model to generate a rating on a 20-point scale. To avoid context effects, all triads were presented separately. Furthermore, the prompts were repeated to address order effects by creating a stimulus list with all six triad permutations of the words *x*,*y* and *z* in different positions (*xy*, *yx*, *xz*, *zx*, *yz*, *zy*).

Prompt: *In this study we want to investigate the degree to which English nouns can be considered related. We will present you with three nouns. Your task is to rate the relatedness of each pair in the triad on a scale from 0 – 20. A rating of 0 means the pair have no possible relation. A rating of 20 means that the pair has the highest possible degree of relation. Evaluate relatedness solely in regard to the meaning of the words, rather than the similarity between other things like letters or rhyme. Think of relatedness in the broad sense. You will now be provided with instructions to complete the task. Please explain the process of each step.*
Step 1: Closely analyse the semantic relatedness of all possible pairs. Step 2: Explain to what degree each pair is related or unrelated. Step 3: On the basis of Step 1 and 2, report the ratings of semantic relatedness for each pair as instructed.
You will now be provided with two examples of relatedness ratings. Example 1: Presented: cold; hot; square Response: cold – hot: 20 cold – square: 0

Now that you have been provided with these practice examples, you are ready to complete the task. The relation of word meaning may be very weak. Even then, assign a rating to each pair. Following is the triad. {*x,y,z*}.

For most prompts, the model successfully formatted the response as instructed, which allowed us to extract the word pair and rating automatically. Occasionally, the model generated a verbose response, which did not fully conform with instructions. The prompt was repeated in those cases until the desired output format was obtained.

Results

Across all model responses, the model adhered to the instructions and ratings were consistently given for values between 0 and 20. Pairwise orders were summed (e.g. $xy = xy + yx$), and aggregated human preferences for pairs *xy*, *xz*, and *yz* were normalized to sum to one. Like human preferences, model ratings were normalized by dividing the similarity value of an option by the total similarity value. Similar to previous studies, we used Pearson correlations to compare human preferences with model predictions.[3] A scatterplot of the data in Figure 1 shows a positive linear relation between human preferences and model predictions. In all cases, the effect sizes were large (Cohen, 2013), $r(298) = .77$, $CI_{95} = [.73, .81]$ for abstract triads, $r(298) = .81$, $CI_{95} = [.79, .84]$ for concrete triads, and $r(298) = .84$, $CI_{95} = [.80, .87]$ for remote triads.

Similarity vs Relatedness

To better understand what might explain the model's success, we investigated the role of instructions and presentation and how these might interact with different ends of the scale (remote vs basic-level) or different kinds of concepts (abstract vs concrete). To do so, we contrasted the notion of relatedness with similarity (Hill, Reichart, & Korhonen, 2016). We also manipulated the nature of the comparisons by presenting the triad options in a pairwise format. This allows us to compare the performance with pairwise similarities from previous top-ranking models more directly. It also allows us to determine how robust the models are using different tasks. This is important as previous studies have shown that human measurements using different methods (e.g. pairwise similarity,

[1] Scripts and data are available at `https://github.com/SimonDeDeyne/cogsci2024`.

[2] We also explore the performance for GPT-3.5. The results were consistently worse, and inspection of the output demonstrated that the model struggled to generate relatedness ratings for weak pairs because it could not find any way they were related.

[3] Note that the p-values are somewhat inflated since the observations are not independent. For this reason, we opted for bootstrapped confidence intervals and p-values. Results with Spearman correlations were highly consistent with the current results, so we focus on Pearson to allow for comparability with previous work.

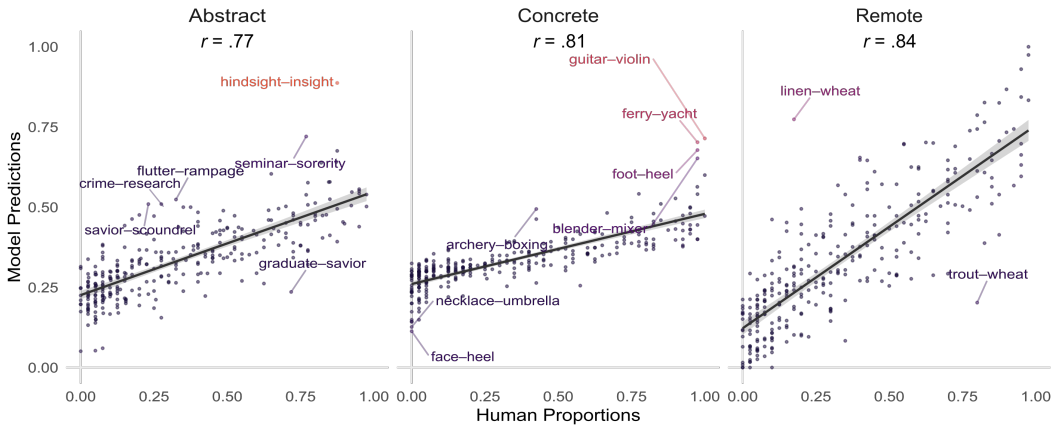

Figure 1: Scatterplots with human (x-axis) and model predictions (y-axis) for abstract, concrete and remote triads. Influential cases with Cooks'D >.025 are labelled in the plots.

triads, sorting) result in highly correlated representations. In contrast, using similar instructions, models like GPT-3 did not converge on similar representations (Suresh, Padua, Mukherjee, & Rogers, 2023).

Beyond comparing different types of comparisons, the current comparison also carefully controls the number of judgments, which allows us to determine whether the performance over at least six aggregated permuted triadic judgments in the previous section might have boosted performance compared to cases where ratings are not repeated as often.

Procedure

In the relatedness condition, the instructions were as follows: *In this study we want to investigate the degree to which English words can be considered related. Words are related if they co-occur in similar situations or have similar meanings. Your task is to rate the relatedness of a word pair on a scale from 0 – 20. A rating of 0 means the pair have no possible relation. A rating of 20 means that the pair has the highest possible degree of relation. Evaluate relatedness solely in regard to the meaning of the words, rather than the similarity between other things like letters or rhyme. Think of relatedness in the broad sense. You will now be provided with instructions to complete the task. Please explain the process of each step. Step 1: Closely analyse the semantic relatedness of the pair. Step 2: Explain to what degree the words are related or unrelated. Step 3: On the basis of Step 1 and 2, report the ratings of semantic relatedness as instructed. You will now be provided with examples of relatedness ratings. Presented: cold – hot Response: cold – hot: 19 Presented: love – affection Response: love – affection: 16 Presented: frog – square Response: frog – square: 0 Presented: dog – bone Response: dog – bone: 13 Presented: car – city Response: car – city: 12 Now that you have been provided with these examples, you are ready to complete the task. The relation of word meaning may be very weak. Even then, assign a rating to each pair. Format your response as follows word1 – word2: rating. Following is the pair.*

The instructions for the similarity condition were generated by changing the wording from relatedness to similarity and updating the examples. The opening was changed to *In this study we want to investigate the degree to which English words can be considered synonymous. Two words are synonyms if they have very similar meanings. Synonyms represent the same type or category of things. Your task is to rate the similarity of a word pair on a scale from 0 – 20.* The examples were changed to: *cold - hot: 4, alligator - crocodile: 18, dog - bone: 2, car - city: 4*.

The instructions were repeated three times for *x-y* and *y-x* pairwise judgements to match the triadic condition. However, the ratings obtained across these iterations were highly similar correlations between .98 and .99. The final model prediction corresponded to the average across the three iterations and words in both orders.

Results

Similar to previous studies, the models were compared by calculating the Pearson correlation between human preferences and model predictions. The results for the relatedness ratings were highly consistent with the results obtained for triadic prompts. For basic-level comparisons, the results were $r(298)$ = .81, $CI_{95} = [.76, .84]$ for abstract triads and $r(298)$ = .74,$CI_{95} = [.69, .79]$ for concrete triads. As before, the correlation was even higher for remote triads $r(298)$ = .85, $CI_{95} = [.81, .88]$. The highly similar effect sizes across triadic and pairwise relatedness instructions suggest that anchoring a third word is unnecessary to obtain the performance reported in the previous section. The results for the similarity instructions were $r(298)$ = .79, $CI_{95} = [.71, .84]$ for abstract, $r(298)$ = .77, $CI_{95} = [.73, .81]$ for concrete, and $r(298)$ = .80, $CI_{95} = [.75, .85]$ for remote triads.

One possibility is that relatedness and similarity make independent contributions across different tasks. To compare the role of relatedness and similarity, we conducted a relative importance analysis using the *relaimpo* R package (Grömping, 2006). The results in Figure 2 show that regardless of the task, relatedness and similarity both contribute to predicting human preferences. For concrete triads, the effect

of similarity was significantly larger than relatedness, and this difference was significant, $\Delta = -.07, CI_{95} = [-0.14, -0.01]$, with similarity explaining a larger portion of the variance compared to relatedness. For remote triads, the difference between relatedness and similarity was also significant, $\Delta = .10, CI_{95} = [0.04, 0.21]$, with relatedness explaining a larger portion of the variance than similarity. This suggests the model differentiates between constructs, and performance interacts with the task.

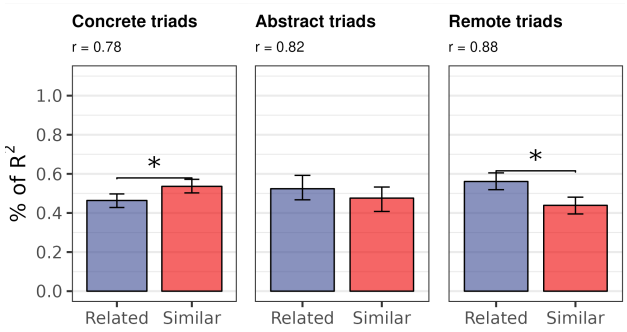

Figure 2: Relative importance with bootstrapped confidence intervals for a regression model predicting human preferences from pairwise relatedness and similarity GPT-4 ratings.

GPT-4 vs SWOW Previous studies have shown that similarity derived from word association data consistently outperforms word embedding models trained on text corpora (e.g., De Deyne et al., 2021; Richie & Bhatia, 2021). In this section, we replicate these results and extend them to investigate how both models capture complementary information. Following the procedure described in De Deyne et al. (2021), we obtained $r(298) = .82, CI_{95} = [.78, .86]$ for abstract basic-level triads, $r(298) = .760$ $CI_{95} = [.71, .80]$ for concrete triads and $r(298) = .76$, $CI_{95} = [.71, .81]$ for remote triads. This shows that depending on what prompting is used, the results for SWOW were better for basic-level triads, whereas GPT-4 provided a better account for remote triads. To investigate whether the contributions between both models made an independent contribution to the prediction of human data, we again performed a relative importance analysis, which partitions both model's contribution to the R-square value. Bootstrap significance tests that compared effect sizes are shown in Figure 3. As can be seen from the figure, both models made comparable contributions, except for remote triad comparisons, where the effect was significantly larger for GPT-4, with the difference between GPT-4 and SWOW, $\Delta = .18, CI_{95} = [.08, .29]$. Moreover, in terms of explanation, a combined model resulted in performance gains in all three conditions.

Discussion

To what degree do large language models capture structure at every scale? To investigate this question, we compared how a large language model, GPT-4, accounted for both remote and basic-level items. Across two evaluations, GPT-4

Figure 3: Relative importance with bootstrapped confidence intervals for a regression model predicting human preferences from pairwise GPT-4 and SWOW relatedness ratings.

captured human behaviour extremely well in both an absolute sense and relatively, compared to previous models with correlations consistently $>.74$ at the basic level and $>.84$ for remote triads. These results were robust and held regardless of whether the model was prompted to perform triadic or pairwise judgments. Despite the different nature of the task, the performance for remote triads was on par with that of basic-level comparisons, which suggests that remote structure is well-represented in humans and Large Language models alike. This refutes strong claims about similarity being too flexible or ill-determined (Goodman, 1972). Here, we tested a more moderate version, which showed that similarity is potentially underconstrained in distributional semantic representations based on language, especially when distributional overlap would be spurious. This idea is put to the test in remote triads. It also suggests that compared to word embeddings (cf De Deyne et al., 2021), perceptual/emotional information is sufficiently encoded when comparing concrete and abstract basic-level words.

Comparison with previous work

Previous work has investigated the prediction using word embedding models such as word2vec. At the basic level, the correlations in De Deyne et al. (2021) with word embeddings trained on a balanced corpus were .64 for concrete triads and .62 for abstract triads, with similar results obtained for GloVe (Pennington, Socher, & Manning, 2014). The same word embeddings were also used to predict remote triads in (De Deyne et al., 2016) and were correlated .52. This suggests that the boost in performance in transformer-based language models is consistent with previous work using challenging pairwise similarity benchmarks such as SimLex (Trott, 2023).

The findings can also be compared with recent work on basic-level similarity in word embeddings that uses a different approach. Richie and Bhatia (2021) conducted a study with basic-level comparisons for categories such as *vehicles*, *birds*, or *sports*. They found that learning category-specific weights for word embedding dimensions in a supervised way substantially improved similarity judgments compared to unweighted embeddings and correlations of .53 for unweighted

embeddings compared to .73 for weighted ones. However, it should be noted that the correlations for the best-performing word embedding model were still lower than those matched to SWOW, and considering the current results for GPT-4, they are likely to be lower to recent transformer-based LLMs as well. Other work has suggested word embeddings might be fundamentally limited in simultaneously capturing similarity and relatedness (Mandera et al., 2017), which could explain the difficulty of capturing a strict notion of similarity (Mandera et al., 2017; Richie & Bhatia, 2021). Our findings suggest LLMs are less affected by this, and both types of representations can be obtained through prompting. Without specific prompting strategies, recent work by Digutsch and Kosinski (2023) showed that GPT-3.5's semantic activation primarily reflects semantic rather than associative similarity. While space does not permit us to go into GPT-3.5, our findings on similarity vs relatedness suggest that prompting can be used to modulate different types of activation, and the flexibility to access these relations might contribute to GPT-4's superior performance for different scales and domains.

Implications

The implications of this work go further than merely showing what others have already argued, namely that LLMs like GPT-4 capture human-like biases, and their prediction represents a jump in performance over previous models, although remain far from perfect in more challenging settings (Han et al., 2024). While demonstrating this remains important at this stage, the current implications, especially those related to the remote triad findings, might explain one aspect of why these models work so well in many cases. Not only are these models trained on vast amounts of data, but the relations between weakly related words are sufficiently constrained in language, despite Goodman (1972)'s concerns, and resemble that of humans. The ability to do so is foundational to other similarity-based approaches that rely on weak similarity. This included methods that rely on anchor words (e.g., *good*, *bad*) to construct any kind of dimensional constructs, notably sentiment or stereotypes (Lewis & Lupyan, 2020). Similarity towards anchor words is likely to be weak, a property it shares with other similarity-based approaches used to predict iconicity ratings, semantic dimensions (i.e. judging age, size, danger of objects) (Grand, Blank, Pereira, & Fedorenko, 2022), metaphor generation, or creativity (e.g., Divergent Association Test, Chen & Ding, 2023).

The consistent performance of prediction similarity at every scale also has implications for semantic theories that posit different mechanisms or processes to judge different types of concepts (e.g. thematic integration for abstract concepts, (Bassok & Medin, 1997), or category-specific weightings for basic-level comparisons, (Richie & Bhatia, 2021)). Relatedness by itself already provided a parsimonious account of various concepts and relations (see Figure 2), except for concrete concepts where similarity was more important. This demonstrates how different types of relatedness (with similarity being more narrowly defined) interact with knowledge domain and distance. Unlike concrete concepts, abstract concepts are supposed to be more relational (Borghi et al., 2017). A comparison between relatedness and similarity prompts was consistent with his view.

Limitations and Future Directions

While the current results suggest that LLM can account for a wide range of human similarity data, these findings do not address cognitive plausibility (especially when considering the massive amount of training data) or questions about how language might encode modal-specific perceptual or affective information (due to the multimodal data, and Reinforcement Learning from Human Feedback part of GPT-4's training). A second practical limitation of using LLMs is that large models can only be approached using an API, which affects replicability. Using simple measures such as a fixed temperature, the results varied only slightly across different runs. However, they may likely vary between different model checkpoints.[4] One of the outputs we have not considered so far is the explanations given by the model as part of the guided triadic prompting. For example, for *clown-rabbit-finger*, the model considered clown and rabbit to be related in "a context of a circus or a magic show where a clown might pull a rabbit out of a hat", clown and finger related through a context where "clown uses his fingers to perform tricks or to apply makeup", and rabbit and finger through a context in which "a person uses their fingers to pet a rabbit". Again, over different runs, the types of explanations tend to vary slightly, but inspection of a large range of outputs suggests systematicity in the types of relations they encode. Future work could consider how this relates to reports by De Deyne et al. (2012), who asked participants to explain their choices, which were subsequently coded using a semantic relation ontology. Most choices were based on thematic information when judging, and to a lesser degree, internal features or category information was used. This finding was most pronounced for remote triads, whereas triadic judgments for more specific artifact and animal domains used a higher proportion of featural and taxonomic relations. This pattern is largely consistent with the relative importance of relatedness and similarity reported in these tasks.

Conclusion

Recent Large Language Models like GPT-4 capture similarity at every scale, from comparisons between weakly related concepts to closely related basic-level comparisons. This ability to capture weak similarity among words with limited additional context suggests another principle that might explain the performance of LLMs across a large range of tasks where this type of similarity matters, pointing towards similarity being sufficiently constrained in language when words occur in vastly different contexts.

[4]Note that the cost of running GPT-4 model predictions factors in our ability to systematically investigate the role of different checkpoints or minor instruction variations.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238–247).

Bassok, M., & Medin, D. L. (1997). Birds of a feather flock together: Similarity judgments with semantically rich stimuli. *Journal of Memory and Language*, *36*(3), 311–336.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165*.

Chen, H., & Ding, N. (2023). Probing the creativity of large language models: Can models produce divergent semantic association? *arXiv preprint arXiv:2310.11158*.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The Small World of Words English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987-1006.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2012). Strong structure in weak semantic similarity: A graph based account. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1464–1469). Austin, TX: Cognitive Science Society.

De Deyne, S., Perfors, A., & Navarro, D. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 1861–1870). Osaka, Japan.

Digutsch, J., & Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in gpt-3 than in humans. *Scientific Reports*, *13*(1), 5035.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437–447). New York: Bobbs-Merrill.

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.

Grömping, U. (2006). Relative Importance for Linear Regression in R. *Journal of Statistical Software*, *17*, 1–27.

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155.

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665-695.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, *4*(10), 1021–1028.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, *36*, 402–407.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). ACL.

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), e13030.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Suresh, S., Padua, L., Mukherjee, K., & Rogers, T. T. (2023). Behavioral estimates of conceptual structure are robust across tasks in humans but not large language models. *arXiv preprint arXiv:2304.02754*.

Trott, S. (2023). Can large language models help augment english psycholinguistic datasets?

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*, 219–247.