scientific reports

OPEN

Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans

Jan Digutsch[1,3] & Michal Kosinski[2]

Modern large language models generate texts that are virtually indistinguishable from those written by humans and achieve near-human performance in comprehension and reasoning tests. Yet, their complexity makes it difficult to explain and predict their functioning. We examined a state-of-the-art language model (GPT-3) using lexical decision tasks widely used to study the structure of semantic memory in humans. The results of four analyses showed that GPT-3's patterns of semantic activation are broadly similar to those observed in humans, showing significantly higher semantic activation in related (e.g., "lime–lemon") word pairs than in other-related (e.g., "sour–lemon") or unrelated (e.g., "tourist–lemon") word pairs. However, there are also significant differences between GPT-3 and humans. GPT-3's semantic activation is better predicted by similarity in words' meaning (i.e., semantic similarity) rather than their co-occurrence in the language (i.e., associative similarity). This suggests that GPT-3's semantic network is organized around word meaning rather than their co-occurrence in text.

Modern large language models (LLMs) employ artificial neural networks that generate texts virtually indistinguishable from those written by humans and achieve near-human performance in comprehension and reasoning tests (e.g.,[1,2]). LLMs are not provided with grammar rules or dictionaries, but are repeatedly presented with a fragment of text (e.g., a Wikipedia article) with one word removed (e.g., "Paris is the capital of _____"), and have to predict the missing word. In the training process, typically involving trillions of trials drawn from huge text corpora, LLMs become skilled language users, spontaneously discovering linguistic rules and word associations.

LLMs' complexity means that it is difficult to explain their functioning and anticipate their future behavior. Both users and creators are often surprised by their emergent properties, both useful (e.g., the ability to translate between languages or write computer code;[3]) and problematic (e.g., gender and racial biases;[1]). It is also unclear whether they are mere stochastic parrots[4] limited to modeling word similarity[5], or if they recognize concepts and could be ascribed with some form of understanding of the meaning of the words they so skillfully use.

The challenge of understanding complex LLMs is not new. The last decades brought significant progress in understanding a much more complex entity capable of generating and comprehending language: the human brain. The methods used to better understand the human brain can be adapted for studying the artificial brain, and there is a growing interest in doing so (e.g.,[6,7]).

To understand and produce language, humans utilize semantic memory that stores information about words and their meaning[8–11]. To unravel the structure of semantic memory, researchers often study the patterns of *semantic activation*, or the phenomenon in which exposure to one word facilitates the processing and retrieval of other words[12]. For example, when asked "What do cows drink?," people tend to answer "milk" instead of "water" (only calves drink milk), revealing that "cow" and "drink" activate "milk" in semantic memory. The research reveals that semantic activation occurs mostly between words that often co-occur in the language (i.e., associative similarity; "wrong–way") and words with overlapping meaning (i.e., semantic similarity; "annoying–irritating";[13,14]). Moreover, while semantic and associative similarity often goes hand in hand, activation readily spreads between the words of similar meaning that rarely co-occur in the language (i.e., purely semantically related words, such as "cow" and "sheep"). This suggests that purely semantically related words are closely connected in the semantic memory, likely through their mutual connections to simple concepts. "Cow" and "sheep," for example, are both linked with "horns," "milk," and "farm"[10,15].

The present research aims to contribute to our understanding of the structure of the semantic memory of humans and LLMs by comparing their patterns of semantic activation. In humans, semantic activation is

[1]Leibniz Research Centre for Working Environment and Human Factors at the Technical University of Dortmund, Dortmund, Germany. [2]Stanford University, Stanford, CA 94305, USA. [3]Institute of Behavioral Science and Technology, University of St. Gallen, St. Gallen, Switzerland. ✉email: jan.digutsch@unisg.ch

typically measured using *semantic priming*, where the exposure to one word (i.e., *prime*) facilitates the processing and retrieval of another word (i.e., *target*). Semantic priming is commonly measured using *lexical decision tasks*[8], where participants are presented with a prime (e.g., "lemon") followed by a real word (e.g., "lime") or a non-word (e.g., "leton"). Participants have to decide, as quickly and accurately as possible, whether the second word is a real word. Their speed and accuracy are interpreted as a proxy for semantic activation. For example, when preceded by "lemon," "lime" is more quickly recognized as a real word than when it is preceded by "dog".

In LLMs, semantic activation can be measured directly from words' distances in the model's lexical space. Early models derived lexical space from word co-occurrences in the training data (e.g.,[16]). They were followed by models capturing words' context in the training data (e.g.,[17,18]). Most recent LLMs employ dynamic lexical space that changes depending on the word's context in a given task (e.g.,[19]). Studies comparing language models' and humans' lexical spaces show that they are increasingly similar as the models become more complex (e.g.,[20–22]).

The present research compares semantic activation patterns between humans and OpenAI's Generative Pre-trained Transformer 3 (GPT-3;[1]), using word pairs typically used in human studies[23]. Analysis 1 compares the semantic activation of GPT-3 and humans across three semantic relatedness conditions and shows that GPT-3's semantic activation patterns broadly mirror those observed in humans. Analyses 2 and 3 compare semantic activation across 12 types of prime-target associations and show that, when compared with humans, GPT-3's lexical space is more strongly organized around semantic (rather than associative) similarity. Finally, analysis 4 compares the relative importance of semantic versus associative activation across five GPT-3 variants as well as three other language models. It shows that the newer the model, the more its semantic space is organized around the semantic similarity.

Methods

Lexical decision tasks (n = 6646) and human participants' responses (n = 768 students from three universities) were obtained from the Semantic Priming Project database ([23]; https://www.montana.edu/attmemlab/spp.html). As it is publicly available and de-identified, its analysis does not constitute human subject research.

Lexical decision tasks consist of a target word (e.g., "lemon") matched with three primes (first-associate, other-associate, and unrelated). Human participants' response times were standardized for each participant separately and then averaged for each word pair. Following Mandera et al.[24], we excluded all non-word trials, erroneous responses, and trials with reaction times deviating more than three standard deviations from the within-person mean.

We used GPT-3's most recent engine aimed at capturing words' similarity ("text-embedding-ada-002"; all settings were left at their default values). It employs a 1536-dimensional semantic space. The location of words or phrases in this space is described by 1536-value-long numerical vectors (i.e., embeddings). In analysis 4, we additionally used four older GPT-3 variants ("text-similarity-[ada/babbage/curie/davinci]-001"), as well as older language models ("xlm-roberta-base", "albert-base-v2";[25,26]).

Semantic activation in GPT-3 was operationalized as the cosine distance between prime and target words' embeddings. Cosine similarity is similar to the correlation coefficient, ranging from – 1 (dissimilar) to 1 (similar). The cosine similarity between "lime" and "lemon," for example, equals 0.35, which is much closer than the similarity between "tourist" and "lemon" (– 0.03). Note that this measure of activation is non-directional: "lime" activates "lemon" as much as "lemon" activates "lime".

We considered an alternative approach, previously used by Misra et al.[19]: presenting GPT-3 with prime words (or sentences containing the prime words) and recording the probability distribution of possible completions. Yet, we believe that this strays too far from the original format of the lexical decision task, which is context-free and does not require participants to complete word sequences. Moreover, the context surrounding the prime becomes a confound, even if it is a mere punctuation sign. The probability of "race" significantly differs among "car race" (log(P) = – 10.27), "car, race" (log(P) = – 9.29), and "car-race" (log(P) = – 6.52).

Separately for humans and GPT-3, semantic activation scores across all prime-target pairs (i.e., first-associate, other-associate, and unrelated) were standardized (mean of 0 and standard deviation of 1) before conducting statistical analysis. To facilitate visual comparisons between humans and GPT-3, the semantic activation displayed was converted to percentiles on the plots.

Results

Analysis 1. We first compare the semantic activation between humans and GPT-3 across three prime word types: The *first-associate prime* (e.g., "lime") is a word to which the target is the most common response (e.g., "Which word comes first into your mind when hearing lemon?"); an *other-associate prime* (e.g., "sour") is a randomly selected word to which the target is a less common response; and an *unrelated prime* (e.g., "tourist") is a randomly selected word to which the target has not appeared as a response (and vice versa).

Figure 1 presents the density distribution of semantic activation for humans (left panel; approximated by the lexical decision task reaction times) and for GPT-3 (right panel; approximated by the cosine similarity between prime and target word embeddings). Like human respondents, GPT-3 shows the highest semantic for first-associate word pairs, followed by other-associates and unrelated word pairs.

However, the differences in mean semantic activation between the priming conditions were larger for GPT-3 than for humans (ANOVA's F(26511) = 1343.70, $p < 0.001$; $\eta^2 = 29.22\%$ or a large effect size versus F(26511) = 104.76, $p < 0.001$; $\eta^2 = 3.12\%$ or a small effect size;[27]). This is to be expected, as GPT-3's semantic activation is measured directly (via cosine similarity), while in humans it is approximated (via semantic priming). Moreover, in contrast with humans, GPT-3 does not suffer from inattention, fatigue, and other response biases. Therefore, GPT-3's results are expected to be more pronounced across all our analyses.

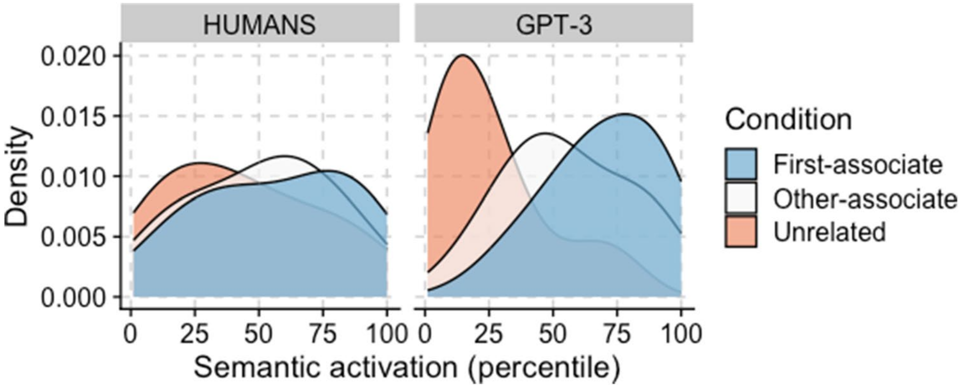

Figure 1. Semantic activation across priming conditions.

Analysis 2. The results of Analysis 1 showed that semantic activation patterns in GPT-3 broadly mirror those observed in humans, but the effect of the priming condition is stronger for GPT-3. Here, we take a closer look by comparing semantic activation across 12 types of prime-target word pairs, following the classification used by the Semantic Priming Project database[23].

The results presented in Fig. 2 show clear differences between semantic activation in humans (red bars) and GPT-3 (blue bars). While for humans semantic activation did not depend strongly on semantic association, clear differences were observed for GPT-3. Its semantic activation was strongest for script (68th), antonyms (68th), categories (63rd), and synonyms (62nd) and weakest for backward and forward phrasal associates (30th and 29th, respectively) and action (28th).

This indicates that in GPT-3 (but not in humans), semantic activation was strongly driven by semantic rather than associative similarity. Antonym, synonym, and category word pairs share many semantic features (e.g., lime and lemon are both sour fruits of similar shape;[14,15]) and relatively rarely co-occur in language. In contrast, forward and backward phrasal associates share little semantic similarity but often co-occur in language.

Analysis 3. Analysis 2 showed an interesting difference between humans and GPT-3: GPT-3's lexical space seems to be organized more strongly around semantic similarity than in humans. Yet, as in Analysis 1, the clearer pattern of GPT-3's results could be driven by the direct approach to measuring its semantic activation. We fur-

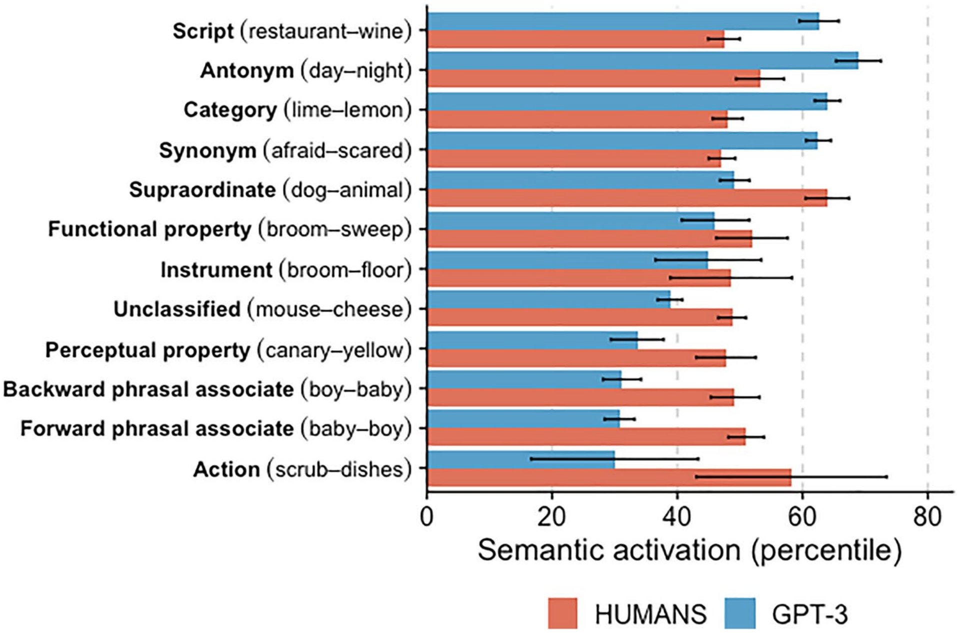

Figure 2. Semantic activation and prime-target association type. Word pairs in brackets are examples. Error bars represent 95% confidence intervals.

ther explore this issue by comparing semantic activation between GPT-3 and humans on the level of individual word pairs.

Table 1 presents word pairs with the largest differences in semantic activation between GPT-3 and humans. For example, "introvert" highly activates "extrovert" in GPT-3 (3.81 SD above the mean), but not in humans (3.06 SD below the mean).

The results confirm the results of Analysis 2. Compared with humans, GPT-3 is particularly prone to be activated by semantically similar word pairs such as antonyms (e.g., "introvert–extrovert" and "understand–misunderstand") and synonyms (e.g., "clumsy–klutz" and "england–britain"). Humans, on the other hand, are relatively more driven by associatively similar words, such as forward phrasal associates (e.g., "mountain–top" and "real–people").

Analysis 4. Analyses 1–3 indicated that, when compared with humans, the semantic space of the most recent engine of GPT-3 was organized around semantic, rather than associative relatedness. Here we compare the relative importance of the associative versus semantic relatedness across several LLMs. Associative relatedness was approximated by averaging the semantic activation for forward and backward phrasal associates. Semantic relatedness was estimated by averaging the semantic activation for antonyms, synonyms, and categories (see "Analysis 2" Section and Fig. 2).

The results, presented in Fig. 3, show that the more recent and more complex the model the more its semantic space is organized around semantic rather than associative relatedness.

Discussion

Our results show that semantic activation in GPT-3 broadly mirrors those observed in humans. GPT-3's semantic activation was significantly higher for first-associated (e.g., "lime–lemon") word pairs than other-associated (e.g., "sour–lemon") or unrelated (e.g., "tourist–lemon") word pairs (Analysis 1). However, the analysis of prime-target association types in Analysis 2 revealed that GPT-3's semantic activation is more strongly driven by the similarity in words' meaning (i.e., semantic similarity) than their co-occurrence (i.e., associative similarity). This effect is stronger in GPT-3 than in humans. In fact, in Analysis 3, the most drastic differences in semantic activation between GPT-3 and humans were observed for synonyms (more similar according to GPT-3) and phrasal associates (more similar according to humans). This suggests that semantic similarity is a stronger predictor of semantic activation in GPT-3 than in humans. Moreover, Analysis 4 reveals that the role of semantic similarity in predicting semantic activation is greater in the more complex and more recent models.

That semantic activation occurs both in humans and GPT-3 is unsurprising. As humans are affected by their semantic activation patterns while generating language, models trained to do the same would benefit from possessing—or simulating—a similar mechanism. It is also possible that the spreading activation (see[28]) is an inherent property of any complex neural network aimed at generating human-like language. To some extent, LLMs may be mirroring (at least on the functional level) semantic structures present in humans. GPT-3's susceptibility

			Z-score	
Prime	**Target**	**Association type**	**Humans**	**GPT-3**
Introvert	Extrovert	Antonym	− 3.06	3.81
Understand	Misunderstand	*Antonym*	− 5.23	0.89
Clumsy	Klutz	Synonym	− 5.65	0.45
Britannica	Britain	Unclassified	− 3.57	2.40
England	Britain	Synonym	− 2.94	2.99
Quality	Characteristic	*Synonym*	− 6.63	− 0.70
Outgoing	Extrovert	Synonym	− 4.28	1.42
Advise	Advice	Unclassified	− 1.64	4.00
Pro	Con	Antonym	− 3.71	1.73
Trait	Characteristic	*Synonym*	− 5.52	− 0.12
Crutch	Leg	Instrument	1.82	− 1.21
Ribs	Broken	Unclassified	1.76	− 1.10
Mountain	Top	Forward phrasal associate	1.84	− 0.98
Brake	Go	Antonym	1.40	− 1.32
Braces	Young	Unclassified	1.22	− 1.48
Approval	Yes	Unclassified	2.00	− 0.64
Clerk	Person	Unclassified	1.69	− 0.92
Real	People	Forward phrasal associate	1.35	− 1.26
Acre	Land	Supraordinate	1.56	− 1.04
Loud	Pain	Unclassified	1.21	− 1.40

Table 1. Largest differences in semantic activation between humans and GPT-3. Association types printed in regular font come from the Semantic Priming Project database; those printed in *italics* were missing and were added by us.

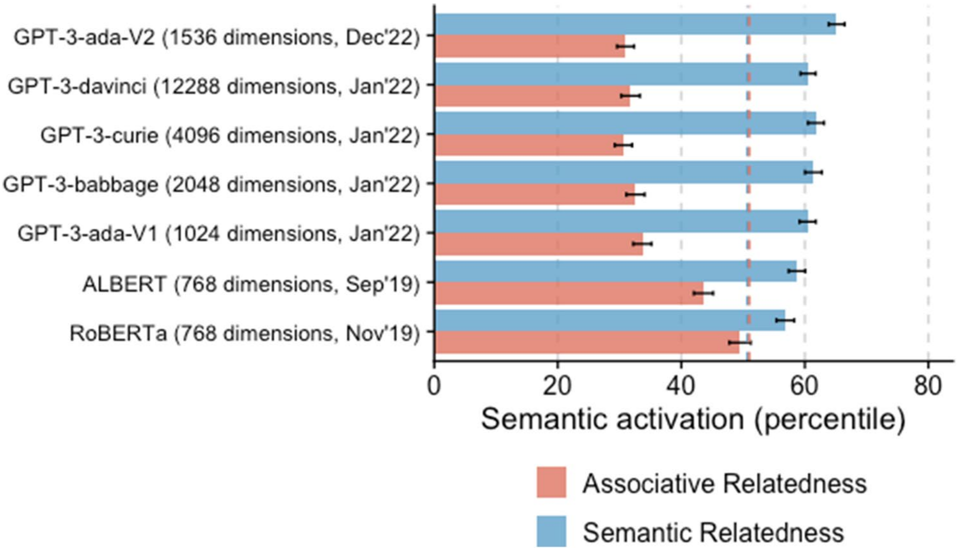

Figure 3. Semantic activation for associatively (red) and semantically (blue) related words. Dotted lines represent semantic activation for the associatively (red) and semantically (blue) related words in humans.

to semantic activation is not the first example of neural networks mirroring human-like psychological processes. Past research has shown, for example, that LLMs mirror human gender and ethnic biases[1], and neural networks trained to process images suffer from human-like optical illusions[29,30]. None of those functions were engineered or anticipated by their developers.

What is surprising, however, is the relatively larger importance of semantic similarity for GPT-3 and the relatively larger importance of associative similarity for humans. It is possible that it is an artifact of the measurement approach used in humans and GPT-3. Semantic priming effects measured using lexical decision tasks in humans are likely affected by processes beyond semantic activation, such as *expectancy* (i.e., an intentional generation of potential completions of a word sequence;[31]) or *semantic matching* (i.e., a retrospective search for the target-prime relationship;[32]). This could be a potential source of noise or bias that is not present in the more direct measure of semantic activation applied to GPT-3 (cosine distance).

Studying LLMs such as GPT-3 could boost our understanding of human language. LLMs are trained to mimic human behavior and could be used as model participants in psycholinguistic studies, enabling researchers to quickly and inexpensively test hypotheses that could be later confirmed in humans (see[33] for a recent example). Unlike humans, LLMs do not suffer from fatigue and lack of motivation, and can respond to thousands of tasks per minute. Moreover, artificial and biological neural networks aimed at processing language may have convergently evolved similar neural structures. As artificial neural structures are easier to study than biological ones, studying LLMs could further the understanding of mechanisms and processes occurring in the human brain. This is not a new idea: The structures of the artificial neural networks trained to process images mirror those observed in the ventral visual pathway[34]. More broadly, the study of convergent evolution has greatly benefited biology, neuroscience, psychology, and many other disciplines (e.g.,[35]). Yet, we should tread carefully: As our results illustrate, LLMs' behaviors are sometimes significantly different from those observed in humans, despite their superficial similarities.

Data availability

Data and code used in the analyses can be found at https://psyarxiv.com/dx5hc.

Received: 23 December 2022; Accepted: 24 March 2023
Published online: 28 March 2023

References

1. Brown, T. B. *et al.* (2020). Language models are few-shot learners. *arXiv* http://arxiv.org/abs/2005.14165 (2020).
2. Van Noorden, R. How language-generation AIs could transform science. *Nature* **605**(7908), 21–21. https://doi.org/10.1038/d41586-022-01191-3 (2022).
3. DeepL. (n.d.). DeepL SE. https://www.DeepL.com/translator
4. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623. https://doi.org/10.1145/3442188.3445922 (2021).
5. Lake, B. M. & Murphy, G. L. Word meaning in minds and machines. *Psychol. Rev.* https://doi.org/10.1037/rev0000297 (2021).
6. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. https://doi.org/10.31234/osf.io/6dfgk (2022).
7. Dasgupta, S., Boratko, M., Mishra, S., Atmakuri, S., Patel, D., Li, X. & McCallum, A. Word2box: Capturing set-theoretic semantics of words using box embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2263–2276. https://doi.org/10.18653/v1/2022.acl-long.161 (2022).

8. Meyer, D. E. & Schvaneveldt, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *J. Exp. Psychol.* **90**(2), 227–234. https://doi.org/10.1037/h0031564 (1971).
9. Katz, J. J. & Fodor, J. A. The structure of a semantic theory. *Language* **39**(2), 170. https://doi.org/10.2307/411200 (1963).
10. Lucas, M. Semantic priming without association: A meta-analytic review. *Psychon. Bull. Rev.* **7**(4), 618–630. https://doi.org/10.3758/BF03212999 (2000).
11. McNamara, T. P. Semantic memory and priming. In *Experimental Psychology. Vol. 4 in I. B. Weiner (Editor-in-chief), Handbook of psychology* 2nd edn (eds Healy, A. F. & Proctor, R. W.) 449–471 (Wiley, 2013).
12. Kumar, A. A. Semantic memory: A review of methods, models, and current challenges. *Psychon. Bull. Rev.* **28**(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x (2021).
13. Holcomb, P. J. & Neville, H. J. Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Lang. Cognit. Process.* **5**(4), 281–312. https://doi.org/10.1080/01690969008407065 (1990).
14. Perea, M. & Rosa, E. The effects of associative and semantic priming in the lexical decision task. *Psychol. Res.* **66**(3), 180–194. https://doi.org/10.1007/s00426-002-0086-5 (2002).
15. Hutchison, K. A. Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychon. Bull. Rev.* **10**(4), 785–813. https://doi.org/10.3758/BF03196544 (2003).
16. Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**(2), 203–208. https://doi.org/10.3758/BF03204766 (1996).
17. Jones, M. N., Kintsch, W. & Mewhort, D. J. K. High-dimensional semantic space accounts of priming. *J. Mem. Lang.* **55**(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003 (2006).
18. Hutchison, K. A., Balota, D. A., Cortese, M. J. & Watson, J. M. Predicting semantic priming at the item level. *Q. J. Exp. Psychol.* **61**(7), 1036–1066. https://doi.org/10.1080/17470210701438111 (2008).
19. Misra, K., Ettinger, A. & Rayz, J. Exploring Bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020* 4625–4635. https://doi.org/10.18653/v1/2020.findings-emnlp.415 (2020).
20. Baroni, M., Dinu, G. & Kruszewski, G. 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)* 238–247. https://doi.org/10.3115/v1/P14-1023 (2014).
21. Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A. & Miliani, M. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Lang. Resour. Eval.* **56**(4), 1269–1313. https://doi.org/10.1007/s10579-021-09575-z (2022).
22. Zhang, L., Wang, L., Yang, J., Qian, P., Wang, X., Qiu, X., Zhang, Z. & Tian, X. Can computers understand words like humans do? Comparable semantic representation in neural and computer systems. *BioRxiv* https://doi.org/10.1101/843896 (2019).
23. Hutchison, K. A. *et al.* The semantic priming project. *Behav. Res. Methods* **45**(4), 1099–1114. https://doi.org/10.3758/s13428-012-0304-z (2013).
24. Mandera, P., Keuleers, E. & Brysbaert, M. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *J. Mem. Lang.* **92**, 57–78. https://doi.org/10.1016/j.jml.2016.04.001 (2017).
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. RoBERTa: A robustly optimized BERT Pretraining approach. *arXiv* https://arxiv.org/abs/1907.11692 (2019)
26. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. ALBERT: A Lite BERT for self-supervised learning of language representations. *arXiv* https://arxiv.org/abs/1909.11942 (2019).
27. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn. (L. Erlbaum Associates, 1988).
28. Collins, A. M. & Loftus, E. F. A spreading-activation theory of semantic processing. *Psychol. Rev.* **82**(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407 (1975).
29. Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M. & Tanaka, K. Illusory motion reproduced by deep neural networks trained for prediction. *Front. Psychol.* **9**, 345. https://doi.org/10.3389/fpsyg.2018.00345 (2018).
30. Benjamin, A., Qiu, C., Zhang, L.-Q., Kording, K. & Stocker, A. Shared visual illusions between humans and artificial neural networks. In *2019 Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/CCN.2019.1299-0 (2019).
31. Becker, C. A. Semantic context effects in visual word recognition: An analysis of semantic strategies. *Mem. Cognit.* **8**(6), 493–512. https://doi.org/10.3758/BF03213769 (1980).
32. Neely, J. H., Keefe, D. E. & Ross, K. L. Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *J. Exp. Psychol. Learn. Mem. Cogn.* **15**(6), 1003–1019. https://doi.org/10.1037/0278-7393.15.6.1003 (1989).
33. Aher, G., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans. https://arxiv.org/abs/2208.10264 (2022).
34. van Dyck, L., Kwitt, R., Denzler, S. & Gruber, W. Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. *Front. Neurosci.* https://doi.org/10.3389/fnins.2021.750639 (2021).
35. Losos, J. Convergence, adaptation, and constraint. *Evolution* **65**(7), 1827–1840. https://doi.org/10.1111/j.1558-5646.2011.01289.x (2011).

Acknowledgements

We would like to thank Gregory Murphy for his extensive and constructive feedback on our manuscript. In addition, we would like to thank Aaron Chibb for his technical support in measuring semantic activation in LLMs.

Author contributions

J.D. and M.K. contributed equally to the planning of the research, the analysis and interpretation of the results, as well as the drafting of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.