LLM-Generated Semantic Networks Predict Semantic Priming Effects on Human Reaction Times in a Word-Recognition Task

David Kinney (kinney@wustl.edu)
Department of Philosophy, Wilson Hall
St. Louis, MO 63130 USA

Abstract

A well-known empirical result in human linguistic processing finds that humans are quicker to correctly recognize a string of letters as word when they are first shown a word that is semantically related to the word they must recognize. This is known as the "semantic priming effect." Since Collins and Loftus (1975), it has been widely theorized that this effect is due to graphical storage of words in memory and a "spreading activation" model of priming. On this theory, words are related to one another in human semantic memory via a graphical structure encoding semantic relationships between words, with participants more likely to quickly recognize a word when they are primed with one that is graphically nearby; the prime word "activates" the node of a participant's semantic memory network representing the prime word and this activation "spreads" to words at nearby nodes. Today, large language models increasingly excel at generating structured data representations, like graphs, when prompted to do so (Ghanem & Cruz, 2024; Dagdelen et al., 2024). In the current paper we investigate whether a language model can be prompted to represent a set of words as a semantic graph, and whether human reaction times in a word recognition task are predicted by the minimum path length between words in such an LLM-generated semantic graph. Using two versions of the Gemini language model, we use a prompting strategy to generate semantic graphs relating all words used in a large semantic priming experiment conducted by Hutchinson et al. (2013), under a variety of different temperatures and settings for the number of maximum output tokens. While we find that all LLM-generated semantic graphs produced during our experiments are such that the minimum path length between two words predicts the reaction time in which a person primed by one word recognizes the other, this effect is most pronounced for graphs generated via a *smaller* version of the model. It is under these conditions, we find, that LLMs produce the dense graphs that are more predictive of human semantic priming effects in lexical decision tasks.

Keywords: semantic priming; language models; cognitive graphs

Introduction

In a lexical decision experiment, participants are shown a word, called the **prime**, followed by the **target**, which is either a second word or a non-word string of letters. Each participant's goal is to determine, as quickly and accurately as possible, whether the target is or is not a word. It is a highly-replicated finding that participants tend to more quickly classify a target as a word when it is preceded by a semantically-related word (Jones & Estes, 2012; Lerner, Bentin, & Shriki, 2014; McNamara, 2005; Meyer & Schvaneveldt, 1971; Neely, 2012). For example, the word *tiger* will be more quickly recognized as a word when it is preceded by *lion* than when it is preceded by *radio*. This is known as the semantic priming effect.

This result is most typically explained by appeal to a cognitive mechanism based on "spreading activation" through semantic networks in memory (Anderson, 1983; Collins & Loftus, 1975; McNamara, 2005; Neely, 2012; Kenett, Levi, Anaki, & Faust, 2017). According to this theory, words in a person's vocabulary are stored in their memory in a graph structure, with a different word at each node and edges representing semantic relationships between those words. For example, *lion* and *feline* might each be nodes in a person's graphically-structured semantic memory, with the *typeOf* relation holding between them. In such a case, a person's semantic memory graph would contain the triple (*lion*, *typeOf*, *feline*). Their semantic memory graph would also contain the triple (*tiger*, *typeOf*, *feline*). Thus, the minimal path length between *lion* and *tiger* in the participant's semantic graph would be no greater than 2. When the participant is shown the word *lion* as a prime, the node corresponding to the word *lion* in their semantic memory graph is activated. That activation then spreads to nearby nodes (e.g., the node corresponding to the word *tiger*), which facilitates processing of those words when they are subsequently shown to a participant. If the target word is sufficiently unrelated to the prime word, (e.g., if the prime is *radio* and the target is *tiger*), then activation will not spread from the prime node to the target node, and no improvement in processing will be observed.

Relatively recent tests of the spreading activation model have largely relied on methods for constructing semantic networks based on human word association data. For example, Kumar et al. (2020) used a dataset compiled by Nelson et al. (2004) consisting of the first English word thought of by each of 150 people after they were prompted by 120 English words. Nelson et al.'s word-association experiment yielded a set 5,018 unique words, which Kumar et al. use to generate a semantic graph in which words are more likely to have a short minimum path length between them to the extent that they are more likely to be associated with each other in the Nelson et al. dataset (see Kennett et al. (2011) for a similar methodology for building graphs representing semantic relationships between Hebrew words). Subsequent results by these authors find correlations between participants' reaction times for a given target-prime pair and the minimum path length between those words in the compiled semantic graph, as predicted by

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

the spreading activation hypothesis for explaining semantic priming results.

Today, large language models (LLMs) have drastically improved our ability to generate on-the-fly, structured representations of natural language inputs (Ghanem & Cruz, 2024; Dagdelen et al., 2024). LLMs are pre-trained on extremely large corpora of natural language text, in order to learn a probability distribution over sub-word tokens that is representative of the statistical patterns of written natural language in general (Radford et al., 2019). They are then fine-tuned using feedback from human evaluators, to improve the quality of text completions and responses (Bai et al., 2022). As such, when they are prompted to output a structured representation of natural language input (e.g., to provide a JSON dictionary representation of a triple in a semantic graph), they are increasingly capable of doing so with a high degree of syntactic and semantic competence.

In light of these breakthroughs, we seek here to explore whether LLMs can, with relatively straightforward prompting, generate graphical representations of the semantic relationships between words such that the minimum path length between target and prime words in said graphs predicts the amount of time it takes a participant to recognize the prime as a word. To this end, we use data from a 512-participant semantic priming experiment conducted by Hutchinson et al. (2013) that generated a total of 4,988 unique words that were used as either a prime or a target. We then used the *Gemini* large language model to build a semantic network out of those 4,988 words through an iterative prompting process, which we repeated under different model parameter conditions and for two different versions of the *Gemini* model. We then measured the degree to which, for each graph, the minimum path length between the nodes representing the target and prime words was predictive of the participant's reaction time in processing the target word.

In every case, we found a significant relationship between minimum path length and reaction time. However, we found that the graph whose minimum path lengths *best* predicted reaction times in a word recognition task (as measured by the log-likelihood of the data according to mixed-effects linear model in which participants are treated as random effects) was generated by a *smaller* version of the *Gemini* model, which we found had a tendency to construct denser, fully-connected semantic graphs, as compared to those generated by the larger version of *Gemini*, which tended to construct sparser graphs with disconnected islands. In what follows, we present and discuss the results of this computational experiment. All code and data used to generate these results is available in a repository at `https://github.com/davidbkinney/llmpriming`.

Experiment

Data

The data used for this computational experiment were collected as part the *Semantic Priming Project* (Hutchison et al., 2013). Specifically, we used data from a 512-participant *lexical decision* semantic priming experiment. On each trial of the experiment, participants were situated 60cm from a computer monitor and shown a fixation cross for 500ms, followed by an upper-case prime word for 150ms, followed by a blank screen for either 50ms or 1,050ms (depending on the "block" of experiments being run), and then finally the target string, which was displayed until it was classified as a non-word or a word or until 3,000ms elapsed. Reaction times (ms) for classifying the target word were recorded. All prime-target pairs came from the aforementioned Nelson et al. (2004) word association dataset. Each participant completed two sessions of 830 trials (415 per block), separated by no more than one week. The result of this experiment on Hutchinson et al.'s part is a 847,724-row data set, with each row consisting of a participant, a session, a block, a prime word, a target word, and a reaction time.

To use this data for our computational experiment, we first removed all rows in which the target word was a non-word. This left 423,889 remaining rows. Together, the set of prime and target words across these remaining rows consisted of 4,988 unique words. It is these words that we sought to embed in a semantic graph, with the goal of using the minimum path length between words to predict human reaction times in a lexical decision task, in keeping with the spreading activation explanation of the data produced by these tasks.

Using an LLM to Build Semantic Networks

We use the *Gemini* family of LLMs to execute an iterative prompting procedure that constructs a semantic graph from an initial input consisting solely of the set of 4,988 unique words extracted from the dataset described above. To illustrate this procedure, let W be the set of 4,988 unique words from the Hutchinson et al. data set. For each $w \in W$, we prompt the language model as follows:

```
You are a top tier algorithm that takes
as input a pair consisting of: 1) a
subject word, and 2) a list of allowed
target words, and returns a list of JSON
dictionaries consisting of all relations
between the subject and the allowed target
words.  The list should be [enclosed in
brackets].  Each JSON dictionary in the
outputted list must contain the following
keys:

-- 'subject': the subject word.
-- 'target': the target word that stands in
a relation to the subject word.
-- 'relation': a one-to-three word phrase
that describes the relationship between the
subject word and the target.

All dictionaries in the list must be
well-formatted JSON. Not all allowed target
```

words need to be related to the subject word. The subject and the target should never be the same word. Do not use any targets not in the list of allowed target words.

Here is the subject word: w.
Here is the list of allowed target words: $W \setminus \{w\}$.

Output only a list. Your output must begin with "[" and end with "]". Each dictionary must begin with a bracket and end with a bracket. Each dictionary must be well-formatted JSON with "subject", "target" and "relation" keys. Non-compliance will result in termination.

Note that the meaning of the word 'target' in the context of this prompt is different from its meaning in the semantic priming paradigm. In the prompt, a 'target' is a possible word for the subject word w to be related to. In the word recognition task, a 'target' is a word to be classified as either a word or a non-word.

The goal of this prompt is to generate a list of triples with the format {"subject":w, "target":w', "relation":r}, where $w' \in W \setminus \{w\}$ and r is a natural-language description, generated by the LLM, of the semantic relationship between w and w'. For example, when w is the word *abdomen* and $W \setminus \{w\} = \{abduct, \ldots, zone\}$, a language model might produce the output:

```
[{"subject": "abdomen",
"target": "stomach",
"relation": "part of"},
{"subject": "abdomen",
"target": "body",
"relation": "part of"}]
```

By repeating this prompt for every $w \in W$, we effectively generate 4,988 lists of JSON dictionaries, with each dictionary specifying a graph triple linking two words in the set of words used in Hutchinson et al.'s semantic priming experiments. This list of triples instantiates a semantic graph representing the relationships between all words in this set.

When prompting a language model in this way, numerous parameters can affect output. Most importantly, the **temperature** parameter affects the amount of noise used when generating tokens, given previously generated tokens. When temperature is 0, the generative language model is effectively deterministic, always outputting what the model takes to be the most likely token, given the previous tokens outputted. As temperature increases, the amount of indeterminacy in the model output increases. For the current study, we completed our full iterative prompting procedure at four different temperatures: 0, .3, .7, and 1. In addition to temperature, we adjust the **maximum number of output tokens** allowed for each LLM call. This allows us to gain some control of the length of LLM outputs and constrain the ability of the model to generate long lists of triples for any one prompt. We ran our full iterative prompting procedure at four different settings for the number of maximum output tokens: 512, 1042, 1536, and 2048.

Finally, we ran our study for two different versions of the *Gemini* model. The first version, gemini-1.5-flash-001, is a smaller model that is designed for high-throughput, low-latency use cases. The second version, gemini-1.5-pro-001, is a larger model with higher latency that is designed for more intensive reasoning tasks. By 'small' and 'large' here were refer to the number of parameters in the statistical model of token co-occurrence learned by either model. As we ran our iterative prompting procedure to generate a graph under all combinations of temperature, maximum output token, and model version parameters, we ended up with 4x4x2=32 semantic graphs relating all unique words used in Hutchinson et al.'s data set. The example model output for $w = abdomen$ provided earlier was generated by gemini-1.5-pro-001 using temperature .3 and 512 maximum output tokens.

Semantic Graph Properties

Before we turn to using these 32 semantic graphs to predict human reaction times in Hutchinson et al.'s lexical decision experiments, we first note some general patterns in the semantic graphs learned under the different parameters defined above. Fig. 1 shows the maximum value of the minimum path length for the graphs produced under each of the parameter combinations listed above, along with whether or not each graph is fully connected. All but one of the graphs produced by the smaller gemini-1.5-flash-001 model are fully connected, and all graphs produced by this smaller model have a maximum value for the minimum path length between any two nodes (excluding completely disconnected nodes) that is either 4 or 5. By contrast, *none* of the graphs produced by the larger gemini-1.5-pro-001 model are fully connected, and all have a maximum value for the minimum path length between any two nodes (excluding completely disconnected nodes) that is between 8 and 10. This suggests that while the smaller model is learning dense, highly-connected semantic networks, the larger model is learning sparser networks that support longer minimum path lengths between nodes. The significance of this difference is confirmed by a paired t-test, which finds a highly significant difference between the two models with respect to the maximum value of the minimum path between any two nodes in the graphs produced by either model ($t = 21.75$, $p = 9.34 \times 10^{-13}$).

In addition to this difference in the graphs produced by the two *Gemini* versions, another key difference between the *processes* by which the two model versions construct generative models concerns hallucinations. In this context, by a "hallucination" we mean a model call in which the resulting output is a triple linking two words where one of those words is not actually an element of the set W of words obtained from the Hutchinson et al. dataset. In other words, a hallu-

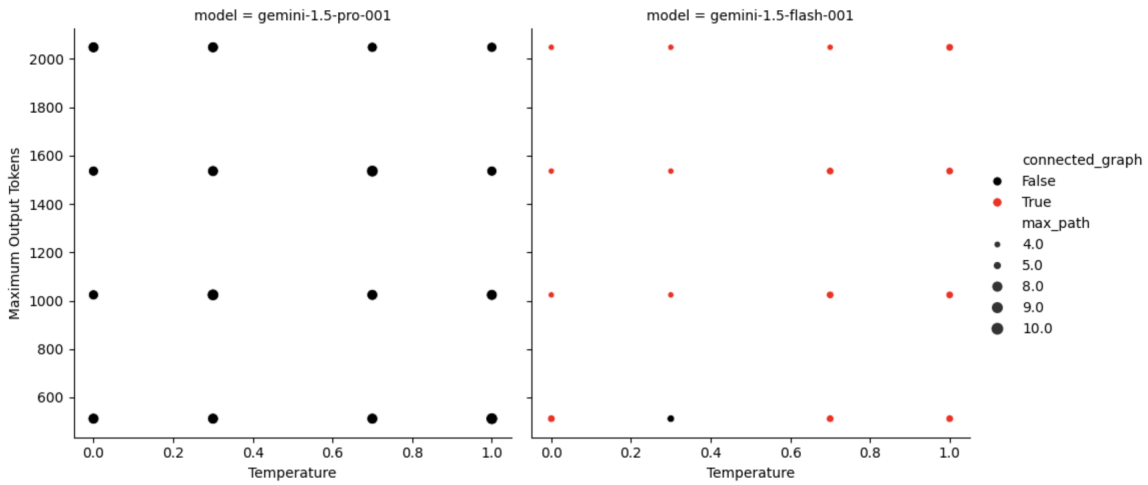

Figure 1: Maximum value of the minimum path length between two nodes in each graph (max_path), and whether the graph is fully connected (connected_graph), organized by model parameters.

cination in this context is a response to the prompt defined above in which the LLM is effectively coming up with new words to include in the semantic graph it is building, despite being asked not to do so. For all 32 processes of building a semantic graph conducted as part of this experiment, we recorded the proportion of triples produced that contained a subject or target word not included in the original set W of unique words in Hutchinson et al.'s data set. For gemini-1.5-flash-001, the mean value of this hallucination rate was .253, while the mean for gemini-1.5-pro-001 was .062. A paired t-test finds that this difference is highly significant ($t = 18.69$, $p = 8.38 \times 10^{-12}$).

Predicting Reaction Times

Our primary motivation for prompting an LLM to construct semantic networks was to assess their value in predicting reaction times in semantic priming experiments. To that end, for each of the 32 semantic graphs generated from the 4,988 unique words included in Hutchinson et al.'s dataset, we went through each of the 423,889 rows in that dataset in which the target is a word, and computed the length of the minimum path between the nodes corresponding the prime word and the target word. For each of the 32 graphs, we fit a mixed-effects linear model to the data, with reaction time as the dependent variable, participant as a random effect (to account for natural variation in participants' reaction times), and the following as fixed effects: minimum path length between the prime and target words in the graph (z-scored), the session in which the prime and target were presented to the participant, and the number of previous tasks that participant had completed. For each of the 32 semantic graphs, we compute the log-likelihood of the observed data set of 423,889 semantic priming experiments, according to this mixed-effects model.

Fig. 2 shows the log-likelihood of the data according each of the 32 mixed-effects models learned according to the procedure defined above, separated by the LLM used to build the

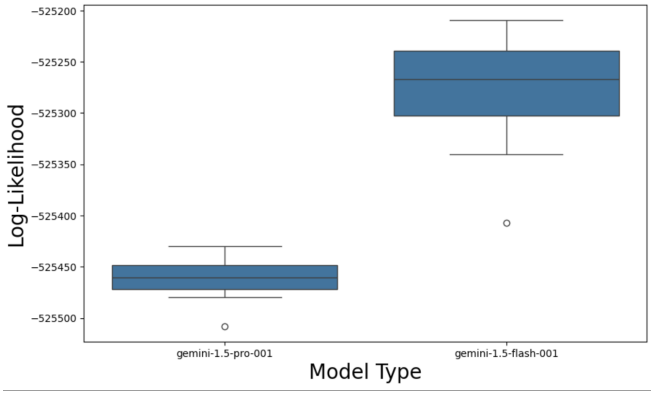

Figure 2: Log-likelihood of semantic priming data according to a mixed-effects model in which minimum path length in a semantic graph is a fixed effect, according to the LLM used to generate the semantic graph.

semantic graph in which the minimum path length variable is calculated. Clearly, the data have consistently higher log-likelihood according to models that use minimum path length in a semantic graph built by gemini-flash-1.5-001 as a fixed effect than they do according to models that use minimum path length in a semantic graph build by gemini-pro-1.5-001 as a fixed effect. That is, the smaller version of *Gemini* appears to build graphs in which the minimum path length between prime and target nodes is a better predictor of reaction time in a word recognition task than graphs built by the larger version of *Gemini*. This is confirmed by a paired t-test, which finds a significant difference between models with respect to the log-likelihood of the data ($t = 13.62$, $p = 7.44 \times 10^{-10}$). More robust statistical analyses, such as bootstrapping analyses, were computationally infeasible given then large size of the data sets used to infer each mixed-effects model, but could be performed using the data provided in our repository.

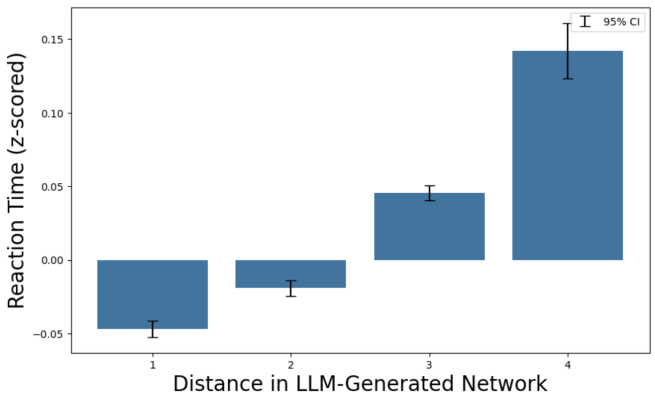

Figure 3: Relationship between minimum path length in the semantic network generated by gemini-1.5-flash-001 with temperature .3 and maximum output token 1024 and the reaction times of participants in the word recognition task (best-performing graph with respect to log-likelihood of Hutchinson et al.'s data).

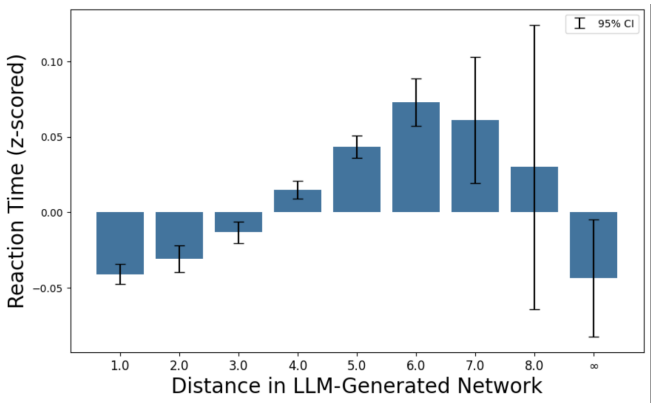

Figure 4: Relationship between minimum path length in the semantic network generated by gemini-1.5-pro-001 with temperature 0 and maximum output token 1536 and the reaction times of participants in the word recognition task (worst-performing graph with respect to log-likelihood of Hutchinson et al.'s data).

The mixed-effects model that confers the greatest log-likelihood onto the data is the model that uses as a fixed effect the minimum path length between the prime and target nodes according to the graph learned by gemini-flash-1.5-001 with a temperature parameter of .3 and 1024 maximum output tokens. In this model, both the minimum distance between nodes corresponding the prime and target words and the number of previous trials a participant had completed were significant predictors of reaction time, with reaction time increasing as function of both the minimum path length between target and prime ($\beta = .044$, $p < .001$) and the number of trials already completed ($\beta = .071$, $p < .001$). See Fig. 3 for a chart showing the relationship between the minimum path length between a prime and a target node in this best-performing LLM-generated semantic network and the reaction times of participants in Hutchinson et al.'s data set.

By contrast, the *worst* performing mixed-effects model by log-likelihood uses as a fixed effect the minimum path length between the prime and target nodes in the semantic network built using gemini-pro-1.5-001 with temperature parameter 0 and maximum output tokens 1536. Here, we see a similarly significant effect on reaction time of both the minimum path length between target and prime ($\beta = .031$, $p < .001$) and the number of trials already completed ($\beta = .071$, $p < .001$). So, even in our worst-performing models, path length in an LLM-generated semantic graph is still a strong predictor of reaction times in the word recognition task. That said, it is clear from Fig. 4 that in this worst-performing model, the relationship between reaction time and minimum path length in the network becomes far noisier, especially at larger minimum path lengths between nodes. Moreover, once we move to cases where the prime and target are not connected at all in the graph, reaction times behave more like they do when the prime and target are graphically very close. This suggests a further instability between reaction times and graphical distance in this worst-performing graph that is not present in the best-performing graph.

Discussion and Conclusion

Our results demonstrate a capability of large language models that, to our knowledge, has yet to be demonstrated: via a relatively straightforward prompting procedure, they can be used to generate semantic networks whose graphical structure predicts semantic priming effects on human performance in a word recognition task. That the minimum path length between nodes in these LLM-generated semantic networks predicts human reaction times in a lexical decision paradigm suggests that the semantic networks the LLM generates are not radically different to the semantic networks that we would generate if confronted with the same set of words.

It is remarkable that the smaller, less powerful, more hallucinatory version of *Gemini* tends to build semantic graphs that have more predictive power, with respect to the relationship between the minimum path length between nodes and reactions times in semantic priming experiments, than the larger, more powerful version. One immediately noticeable difference between the graphs produced by the two model versions is the difference in density between the two groups of graphs. See Fig. 5 for a visual comparison of the two graphs: clearly the best-performing graph is far denser than the worst-performing one. This suggests that, if the spreading activation model of semantic priming is correct, then the networks that activations spread over should have a density closer to those generated by gemini-1.5.flash-001.

More speculatively, one might argue that these results support the conclusion that in humans, the process of building semantic networks from one's existing vocabulary is more likely to be the implemented via a less highly parameter-

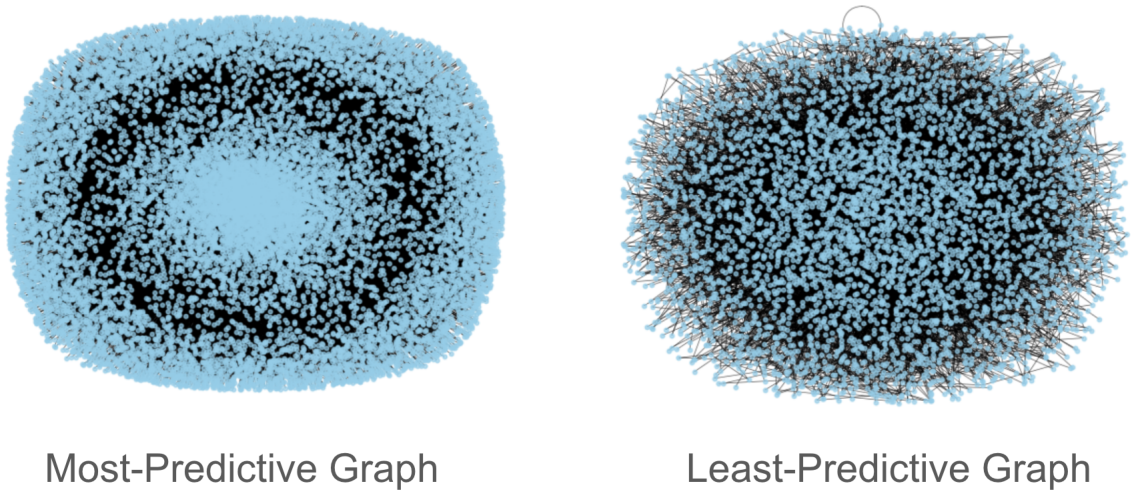

Figure 5: Comparison of the density of the best-performing and worst-performing graphs with respect to using the minimum path length between the nodes corresponding to two words to predict human reaction times in the word recognition task.

ized statistical model that prioritizes speed over deft reasoning. Such an argument would require at least some willingness to analogize between the performance of generative language models and human reasoning, but if we allow ourselves this analogy, then our results here do suggest that when we store semantic meanings graphically in our memory, this storage process is not a highly deliberative process of reasoning but instead a more promiscuous binding together of semantic concepts, leading to a dense thicket of semantic connections with no island or gaps. Put somewhat more simply, our results here suggest that building semantic networks may be a relatively low-effort cognitive endeavor. Moreover, the high rate of hallucinations among the better-performing models suggests that a certain amount of creativity and relaxation of constraints may be part of the natural process of generating semantic networks.

One problem for this line of argument is that if semantic networks are learned by humans at all, then they are not learned through responding to a prompt asking a person to identify relations between a large list of words. Instead, a person's semantic network is learned dynamically through their encounters with natural language. In response this worry, we clarify that, in our view, the semantic networks generated through our prompting strategy are already latently encoded in the weights of each LLM; our particular prompting strategy amounts to an explicit rendering of what is already latently stored. These model weights are learned through a dynamic process of pre-training on large text corpora which, while decidedly different from human language learning, is arguably closer to people's actual encounters with natural language than the prompting we use to extract and make explicit the latently stored semantic graph. That the semantic networks stored latently in language models are more human-like when those models have *fewer* parameters may still have important implications for how such semantic networks are learned in humans.

At the same time, there are some significant limits to the results presented here. First, we have only run our experiments for a single model family: the *Gemini* family. In future work, we hope to extend our analysis to other model families (e.g., GPT, Claude, or Deepseek). If we were to see a similar pattern of smaller models generating denser graphs wherein the minimal path length between nodes is more predictive of a priming effect on reaction times, then this would provide stronger evidence for the cognitive hypothesis that information-processors generally trade careful reasoning for speed when generating semantic networks. It will also be necessary to compare the predictive accuracy off LLM-generated and non-LLM generated (e.g., Kumar et al., 2020) semantic graph models with respect to human reaction times in the lexical decision task. Second, we hope in future work to test our results on a wider range of word lists, including words in languages other than English. This would allow us to further test the universality and cognitive significance of our results herein. That is, if we were to find a similar pattern in other languages where denser networks learned by smaller models are more predictive of human semantic priming effects in a word recognition task, then it would lend further evidence to the claim that generating semantic networks is a process guided by frugality.

In conclusion, when LLMs are prompted to explicitly render latent representations of semantic networks, we find that the semantic graphs that they build seem to provide a high-fidelity reconstruction of how many of us organize concepts in our own minds. This capacity of LLMs warrants further investigation to determine the implications of this finding for understanding our own semantic storage mechanisms.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261–295.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … others (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., … Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418.

Ghanem, H., & Cruz, C. (2024). Fine-tuning llms or zero/few-shot prompting for knowledge graph construction? In *French regional conference on complex systems.*

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior research methods*, *45*, 1099–1114.

Jones, L. L., & Estes, Z. (2012). Lexical priming: Associative, semantic, and thematic influences on word recognition. In *Visual word recognition volume 2* (pp. 44–72). Psychology Press.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the hebrew mental lexicon. *PloS one*, *6*(8), e23912.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470.

Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2261.

Lerner, I., Bentin, S., & Shriki, O. (2014). Integrating the automatic and the controlled: strategies in semantic priming in an attractor network with latching dynamics. *Cognitive science*, *38*(8), 1562–1603.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading*, 264–336.

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, & Computers*, *36*(3), 402–407.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.