Semantic Dynamics in LLMs: A Comparison Between Spreading Activation and Vector-Based Similarity

Anna Ferrari, University of Trento

May 11, 2025

Abstract :

Large Language Models (LLMs) have achieved human-like complexity and specificity in natural language processing, challenging traditional distinctions between machine and human language capabilities. However, it remains unclear whether(and, potentially, to what extent) the semantic network of LLMs and its processes resemble those of human cognition. This study investigates the underlying semantic organization of LLMs by comparing three approaches to semantic similarity: spreading activation patterns in LLM-generated semantic networks, human-like similarity judgments, and vector-based similarity measurements. Using a semantic network constructed with Mistral AI, we simulate spreading activation processes and compare the results with both LLM-generated semantic relatedness judgments and cosine similarity scores from word embeddings. Our findings reveal that cosine similarity correlates more strongly with LLM-generated similarity judgments (r=0.467) than spreading activation values do (r=0.37), suggesting that LLMs rely more on static geometric proximity in embedding space than on dynamic associative processes characteristic of human cognition. While embedding models are trained on co-occurrence data, they transform this information into distributed semantic representations that capture second-order relationships between concepts. This study offers insights into the semantic mechanisms of LLMs and highlights fundamental differences between computational and human approaches to semantic representation, contributing to our understanding of both artificial and human language processing.

1 Introduction

The performance of modern Large Language Models (LLMs) in natural language processing (NLP), both in text comprehension and generation, has improved significantly in recent years. Despite these advances, LLMs are often regarded as "black-box" systems due to their lack of interpretability. As a result, it remains unclear to what extent their approach to language processing aligns with that of humans. This opacity can pose significant limitations, particularly in contexts where precision and reliability are critical.[1] This paper aims to reconstruct and describe a framework to understand how LLMs retrieve and represent semantic knowledge. By comparing their performance across semantic similarity tasks, including spreading activation, vector-based similarity, and human-like judgment simulations, we aim to evaluate the extent to which LLMs reflect cognitively plausible mechanisms of semantic memory organization.

1.1 Similarity tasks

In particular, three different approaches to semantic similarity will be investigated, namely

- Semantic priming through Spreading activation, a cognitive theory that models the functioning of semantic memory (responsible for meaning, understanding, and concept relationships). According to this theory, knowledge is organized as a net of interconnected concepts; if one is activated, activation spreads automatically to nearby nodes (see Figure 1). “Semantic priming”, which is what can be observed in real-world applications, is the phenomenon in which exposure to one word facilitates the processing and retrieval of other words, that have been “activated” due to a semantic similarity.[2]
- Human-like judgements of level of association in word-pairs, in a scale 1-7.
- Cosine similarity in word embeddings, using word2vec pretrained model. The word embedding technique employed is the so-called Continuous-Bag-Of-Words (CBOW), a neural network approach that provides probability of how a likely other words will be in the context of target word and puts these weights into a word vector representation. [3] Cosine Similarity was calculated as:

$$\text{Cos\_Sim}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

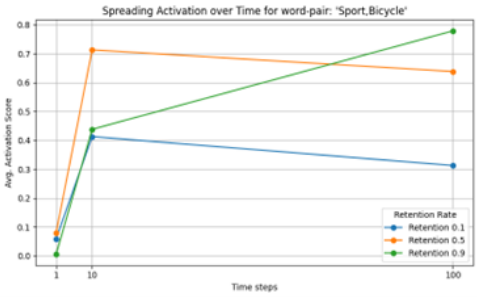

Figure 1. Visualization of Spreading Activation with different parameters

The present research compares spreading activation patterns in

[1] Marcus, G. (2020, February 14). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv.org. https://arxiv.org/abs/2002.06177

[2] Digutsch, J., Kosinski, M. (2023b). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-32248-6

[3] Gerth, T. (2021). A Comparison of Word Embedding Techniques for Similarity Analysis. Computer Science and Computer Engineering Undergraduate Honors Theses Retrieved from https://scholarworks.uark.edu/csceuht/85

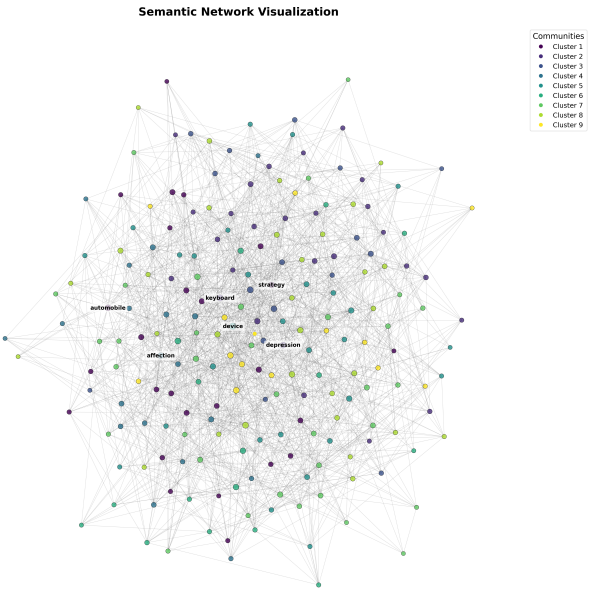

Figure 2. LLM-generated graph with clusters

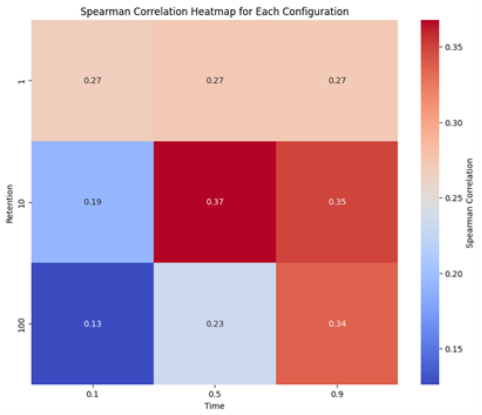

Figure 3. Heatmap of correlation between Spreading activation results and Likert Scores

LLM-generated semantic networks with human-like similarity judgments, which are typically based on semantic relatedness. It also investigates whether there is a correspondence between these activation-based patterns and vector-based similarity measures—specifically, cosine similarity computed using Word2Vec, which captures word meaning based on co-occurrence and distributional context.

2 Methodology and results

To model the internal semantic structure of a large language model, we constructed a semantic network using Mistral AI. Firstly, we selected a set of input words and for each one the LLM was prompted to generate a list of the most semantically related concepts. The responses were used to construct an undirected graph where each node represents a concept (max 2 words) and each edge represents an associative link (deduced from the model's output). The graph obtained contained 210 nodes and 1500 edges, and a density of 0.067. In Figure 2 different clusters of words are highlighted with different colors. This graph was used to simulate spreading activation patterns in a semantic network. Furthermore, to assess the sense of semantic similarity of the model, Mistral AI model was presented with randomly generated pairs of words present in the generated network and asked to rate their similarity on a Likert scale (1 = 'not similar at all', 7 = 'highly similar'). Ratings were treated as proxies for human-like semantic relatedness judgments and used as a reference for comparison with both spreading activation results and vector-based cosine similarity scores. Lastly, vector-based cosine similarity values were calculated using the Word2vec pre-trained model "glove-twitter-50". Words that were not available in the Word2vec library, but appeared among randomly selected pairs, were deleted from the analysis. To model semantic activation between word pairs, we simulated a spreading activation process using the spreadpy package in Python. For each word in a randomly generated pair, activation was initiated from the other word, and the resulting activation levels were computed. The final activation score assigned to each pair was the average of the two directional activation levels. In order to calibrate the two main hyperparameters of the model (retention rate and time steps) we tested several combinations: retention rates of 0.1, 0.5, and 0.9, and time steps of 1, 10, and 100. Each resulting configuration was evaluated by computing the Pearson correlation between the simulated activation scores and the LLM's semantic relatedness ratings collected using a Likert scale. The optimal configuration (retention rate = 0.5, time = 10) was selected based on both the strength and the stability of the correlation with human-like ratings (see Figure 3). Specifically, this configuration consistently produced higher correlations (avg= 0.37), and the difference in Fisher-transformed Pearson's r values between this setup and the second-best configuration was statistically significant ($p = 0.0039$). This choice reflects a balanced modelling assumption: a moderate retention rate suggests that both immediate and sustained activation are relevant, while a time window of 10 steps avoids capturing purely automatic (very short time) or excessively diffuse (very long time) semantic activation. Therefore, all subsequent analyses are based on activation scores computed using this parameter setting. Secondly, the averaged activation scores obtained from the selected configuration were compared to the cosine similarity of word embeddings. The resulting correlation was 0.327. Although this configuration still yields the highest correlation with the compared parameter (in this case, cosine similarity), thus supporting the hyperparameter tuning choices, the strength of this correlation is notably lower than that observed with Likert scores. Lastly, we compared LLM generated Likert scores of semantic similarities with cosine similarity of word embeddings. The correlation between the two values was by wide margin higher than the previous two comparison, with a Sperman's r value of 0.467. (see Figure 4)

3 Discussion and conclusions

Our analysis demonstrates that cosine similarity correlates more strongly with Likert scores (generated by a large language model (LLM) simulating human judgments of semantic relatedness) than spreading activation values do. This result carries both method-

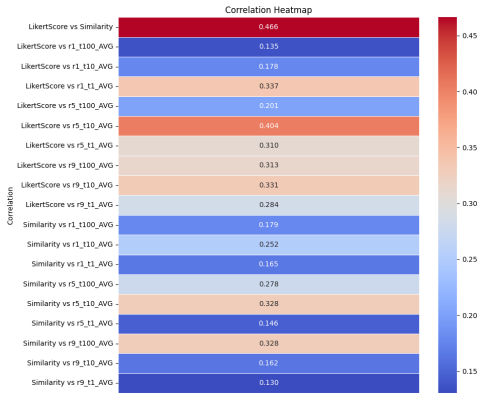

Figure 4. Correlation heatmap of all combinations

ological and theoretical implications. On a methodological level, cosine similarity appears more effective at capturing surface-level, context-independent semantic similarity. Likert-style judgments (especially those produced by LLMs) likely reflect this type of similarity: a direct assessment of how semantically close two words are, without the influence of dynamic contextual or associative processes. In contrast, spreading activation simulates cognitive mechanisms of memory retrieval, modeling how activation diffuses through a network based on structure and time. It reflects deeper, associative, and topologically mediated relationships, such as those found in human semantic memory, making it better suited to psychological modelling than to mirroring the internal logic of LLMs.

These findings are aligned with those of Digutsch and Kosinski (2023), who showed that GPT-3's semantic activation is primarily driven by semantic similarity (e.g., synonyms, antonyms, taxonomic categories) rather than associative similarity (e.g., forward/backward phrasal associates). Their study revealed that GPT-3 is more sensitive to overlap in meaning than to word co-occurrence patterns, which are more predictive of human responses in tasks involving semantic priming. Our results reinforce this distinction: cosine similarity (which measures geometric closeness in embedding space) aligns well with LLM-generated judgments, while spreading activation, despite being cognitively plausible, does not.[4]

Initally, there seems to be a contradiction: Word2Vec, and similar embedding models, are trained using co-occurrence statistics, yet cosine similarity derived from these embeddings appears to better reflect semantic similarity rather than associative (co-occurrence-based) similarity. This paradox resolves when we understand the transformative power of embedding spaces. While Word2Vec uses co-occurrence as its raw training data, it maps this information into a high-dimensional geometric space that reveals second-order relationships. Words that never directly appear together in text can still end up with similar vectors if they share contextual patterns. Consider "cat" and "dog" - they may rarely co-occur directly, but both frequently appear near words like "pet," "fur," and "tail." The embedding algorithm recognizes these shared contextual environments and positions these conceptually similar words near each other in the vector space, despite their lack of direct co-occurrence. Through this mathematical transformation, embedding models transcend raw co-occurrence statistics to capture distributional semantics. The resulting cosine similarity between word vectors measures this deeper semantic relatedness, which explains why it effectively predicts language model behavior. This geometric representation of meaning differs fundamentally from the associative chains used in spreading activation models of human cognition, highlighting a key distinction between computational and psychological approaches to semantic representation. A significant limitation of this study must be taken into consideration: the stronger correlation between cosine similarity and LLM-generated Likert ratings might be explained by a form of "model consistency" rather than reflecting a genuine property of semantic organization. Since the Likert ratings of word pair similarities were generated by an LLM that likely relies on similar vector-based representations to those used in calculating cosine similarity, the observed correlation could be circular. In other words, we may be measuring the consistency of the model with itself rather than revealing an independent property of semantic processing. This potential confound suggests caution in interpreting these findings and indicates the need for future work incorporating human-generated similarity judgments as an external reference point. In conclusion, while human semantic processing relies on heterogeneous connections between concepts that vary in type and strength, with activation flowing dynamically through associative pathways modulated by context, attention, and emotional states, LLM semantic spaces represent relationships through uniform distance metrics in a fixed embedding structure, where concepts with similar distributional patterns converge regardless of direct co-occurrence. This framework helps explain our empirical finding that cosine similarity correlates more strongly with LLM judgments (r=0.467) than spreading activation values (r=0.37), suggesting that although LLMs can approximate human language output, they do so through fundamentally different semantic mechanisms. LLMs fundamentally rely on static geometric proximity rather than the dynamic associative processes characteristic of human cognition.

[4] 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). https://doi.org/10.1038/s41598-023-32248-6

Bibliography

- Digutsch, J., Kosinski, M. (2023b). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-32248-6
- Gerth, T. (2021). A Comparison of Word Embedding Techniques for Similarity Analysis. Computer Science and Computer Engineering Undergraduate Honors Theses Retrieved from https://scholarworks.uark.edu/csceuht/85
- Marcus, G. (2020, February 14). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv.org. https://arxiv.org/abs/2002.06177
- Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., Cukur, T. (2018). Semantic structure and interpretability of word embeddings. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(10), 1769–1779. https://doi.org/10.1109/taslp.2018.2837384
- Collins, A. M., Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82(6), 407–428. https://doi.org/10.1037/0033-295x.82.6.407