Locating Semantic Understanding: A Study of Semantically Similar Sentences in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown increasing capability in natural language processing tasks, yet how they capture and represent high-level semantic concepts remains underexplored. This study investigates the semantic evolution within LLMs as they encode pragmatic meanings across model layers. We extract and analyze hidden state representations for sentence pair data with varying degrees of semantic similarity. Utilizing various metrics and visualizations, we assess the evolution of semantic meaning across model layers qualitatively and quantitatively. Our results indicate that semantic complexity, sentence structure, and model depth impact when pragmatic meaning is encoded, with local semantics processed in the shallow layers and full sentence comprehension occurring in the deeper layers.

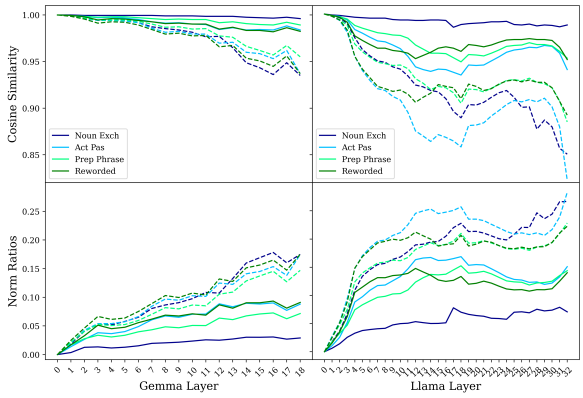

Figure 1: Semantically similar sentences produce highly similar last-token embeddings, particularly in deeper layers. Cosine similarity and norm ratio trends for the last-token embedding are shown for Gemma (left) and Llama (right) across layers. Solid lines represent semantically similar pairs, while dashed lines indicate baseline pairs

1 Introduction

In recent years, transformer-based large language models (LLMs) have enjoyed remarkable success in natural language processing, excelling at tasks such as text generation and comprehension (Vaswani et al., 2017). Despite their widespread deployment, the inner workings of LLMs remain largely opaque, raising concerns about their reliability, safety, and potential biases (Bengio et al., 2023; ACM FAccT, 2023; CAIS, 2023; Anwar et al., 2024; Raiaan et al., 2024). As these models become integral to critical applications, understanding how they process and represent semantic meaning has become an increasingly urgent research priority. While significant strides have been made in interpreting the contextual representations of BERT-like models, particularly regarding their treatment of linguistic information in intermediate layers (Jawahar et al., 2019; Ethayarajh, 2019; Tripodi, 2021, Turton et al., 2020), less is understood about how semantic meaning evolves across layers in generative LLMs, due to the challenges posed by their scale and complex architecture (Zhang et al., 2023; Berglund et al., 2023; Gonen et al., 2024).

Drawing inspiration from Berglund et al. (2023)'s discovery of the 'reverse curse,' where LLMs struggle to generalize inferences in the reverse direction, this work investigates how LLMs internalize pragmatic semantics by examining how they process sentences that share similar semantic meanings but differ in structure and wording. Using pairs of sentences with varying degrees of semantic similarity, we inspect and analyze changes in embedding vectors across model layers to identify where semantic recognition occurs. Our evaluations include measurements of cosine similarity, norm ratio (defined in Section 2), and Kullback–Leibler (KL) divergence. Additionally, we apply multiple representation interpretability techniques, such as probing and principal component analysis, to provide a comprehensive analysis of how pragmatic meaning is encoded across layers.

This study extends previous work on lexical se-

mantics in LLMs by focusing on the emergence of pragmatic meaning, rather than solely the geometric representation of individual concepts. By pinpointing the layers where this meaning emerges, we aim to enhance the transparency and interpretability of LLMs, facilitating the development of more reliable and contextually aware models for real-world applications. In the following sections, we detail the construction of our dataset and experimental design in Section 2, and present the results of our analysis in Section 3. Reviews of related work can be found in Appendix A. Our key findings are summarized as follows:

- Semantic complexity and sentence structure influence how models encode pragmatic meaning, with models sometimes failing to fully reconcile differences in complex reworded sentences
- Sentences with high semantic similarity and sentence structure produce highly similar last-token embeddings (see Fig. 1).
- LLMs tend to encode information locally in shallow layers, with full sentence comprehension emerging in deeper layers.

2 Experimental Design

Datasets We generate four types of sentence pairs using GPT-4o (OpenAI, 2024), ensuring each sentence pair conveys the same or highly similar semantic meaning. The first type, **Noun Exchange** (`Noun Exch`), involves swapping the positions of two common names without altering meanings. The second type, **Active-Passive Voice** (`Act Pas`), pairs an active voice sentence with its passive voice counterpart. The third, **Prepositional Phrase** (`Prep Phrase`), swaps the position of the prepositional phrase within the sentence. Lastly, **Reworded Sentences** (`Reworded`) are sentences that have been rephrased to retain the same meaning, often with notable differences in wording. We manually create initial examples of each type and prompt GPT-4o to generate 500 pairs per category. Intuitively, we anticipate that `Noun Exch` pairs will exhibit the highest semantic similarity, while `Reworded` pairs will show the lowest. Further details and examples for each type are provided in Appendix B.

Prompting Configuration Because the last-token embedding is used for next-token prediction in the LLM and is central to our analysis, it is crucial that the next token generated is meaningful. If a period ('.') is used as the last token, the model will likely predict a space character as the next token, offering minimal insight. To address this, we append specific terms to the original sentences to ensure that the next token generated is substantive and leverages the full meaning of the preceding sentence. Through iterative testing, we found that appending "They" to sentences of the `Noun Exch` type and "And" to the other three sentence types yields optimal results. Furthermore, we experimented with using a consistent last token across all sentence types by appending "Generate an alternative version of the sentence that conveys the same idea but uses different words:" to the input for Llama, which led to smaller variations in the resulting embeddings. This strategy was not applied to Gemma due to its smaller size which resulted in difficulties producing coherent completions with the prompt.

Language Models We conduct our analysis using open-weight models Gemma 2B with 18 layers and Llama 8B with 32 layers (Team, 2024; AI@Meta, 2024), accessed through the HuggingFace `transformer` library (Wolf et al., 2019). The varying parameter sizes allow us to examine how the ability to process semantically similar sentence pairs scales with model size and depth, and whether greater model depth leads to enhanced capacity to capture nuanced semantic meanings.

Evaluation Metrics We use **cosine similarity** to quantify the similarity between embedding vectors at the final token position.[1] Additionally, we introduce a metric termed **norm ratio** to measure the difference between two embedding vectors. For two n-dimensional embedding vectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, norm ratio is defined as

$$\frac{\|\boldsymbol{u}-\boldsymbol{v}\|}{\|\boldsymbol{u}\|+\|\boldsymbol{v}\|}=\frac{\sqrt{\sum_{i=1}^{n}(u_i-v_i)^2}}{\sqrt{\sum_{i=1}^{n}u_i^2}+\sqrt{\sum_{i=1}^{n}v_i^2}},$$

where u_i and v_i represent the ith components of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Both metrics are applied to the last-token embedding for each sentence in a sentence pair across all intermediate and final layers. To further compare the semantic similarity across different sentence subtypes, we compute the **Kullback-Leibler (KL) divergence** of the logits derived from last-token embeddings in the final model layer.

[1] We refer to this as the last-token embeddings for the rest of the paper for brevity.

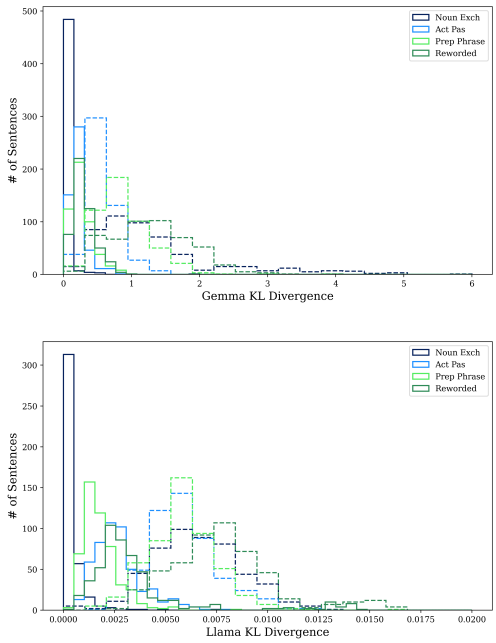

Figure 2: KL divergence for final layer last-token logits in Gemma (left) and Llama (right) shows `Noun Exch` pairs have the highest similarity in semantic meaning and sentence structure. Solid lines represent matched pairs with lower KL divergence, while dashed lines show unmatched pairs as a baseline

Principal Component Analysis We apply PCA to two sets of embeddings: the last-token embeddings for each sentence from a sentence pair, and the difference vectors between the last-token embeddings from 500 matched and 500 randomly sampled unmatched sentence pairs. For each sentence type in each set of embeddings, we project the embeddings on the average pairwise differences and sums of the last-token embeddings from matched sentence pairs: avg sum/diff $= \frac{1}{500}\sum_{i=0}^{500}\left(u^i \pm v^i\right)$. We visualize and examine the evolution of these embeddings across model depth.

Probing We train logistic regression probes to distinguish the embedding pairs of matched and unmatched sentence pairs across model layers. The dataset consists of the log of the norm of the difference vectors between the last-token embeddings of 500 matched pairs (labeled 1) and unmatched pairs (labeled 0), with an 80/20 split for training and testing. We used the `Logistic Regression` function from the `Scikit0Learn`[2] library with L2 regularization. Regularization parameters and probability thresholds are tuned to optimize test performance.

[2] https://scikit-learn.org/stable/

3 Results and Discussion

Trend Across Layers Figure 1 shows the evolution of average cosine similarity and norm ratio compared with 500 randomly sampled baseline pairs for each sentence type. In both LLMs, cosine similarity remains high for matched pairs, with a slight overall decrease, while the norm ratios increases. The higher cosine similarity for matched pairs compared with baseline pairs confirms that the metrics effectively capture semantic similarity. Among the sentence types, `Noun Exch` pairs consistently exhibit the highest cosine similarity, while `Act Pas` and `Reworded` pairs show lower cosine similarity, suggesting that the models handle syntactic variations and paraphrasing differently. The initial decline followed by a slight increase in cosine similarity in Llama indicates that the model effectively learns local semantic meaning in the shallow layers, while the deeper layers (around 60%) further reconcile the semantic meaning of the entire sentence between the sentence pairs.

Figure 2 presents the KL divergence distribution for the final layer's last token logits. As expected, both models exhibit the lowest KL divergence for `Noun Exch` pairs among all sentence types. The other sentence types display similar KL divergence distributions. Notably, the KL divergence for matched sentence pairs is lower than for randomly sampled unmatched pairs, suggesting that the last-token embedding effectively captures the semantic meaning of the entire sentence.

PCA Additional PCA visualizations for `Noun Exch`, `Reworded`, and the difference vectors between matched and unmatched pairs are provided in Appendix C. In both models, `Noun Exch` embeddings show minimal separation across layers, while `Reworded` embeddings exhibit moderate initial separation that remains stable throughout. In contrast, as illustrated in Figure 3, `Act Pas` and `Reworded` embeddings begin with some separation but gradually converge across layers, though not to the same extent as `Noun Exch` embeddings.

These trends suggest that the models progressively reconcile sentence structures, especially after about 60% of the layers. Semantic complexity appears to influence the degree of separation and convergence, with `Noun Exch` pairs showing the highest embedding similarity, followed by `Prep Phrase`, while `Act Pas` and `Reworded` show the lowest. We note that this convergence may be driven by shared structural features rather than

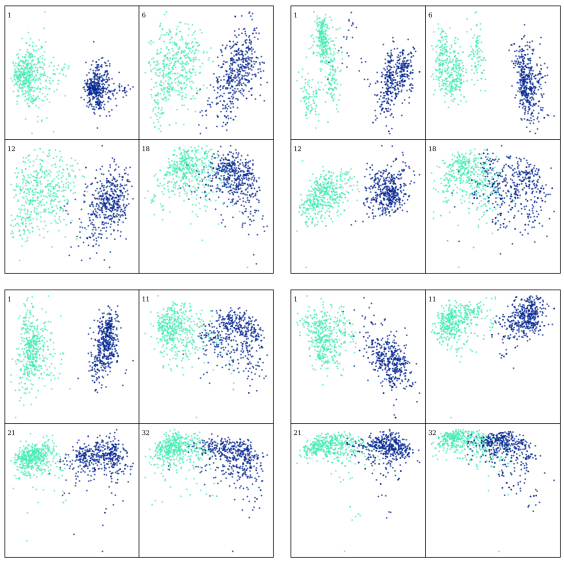

Figure 3: PCA visualizations reveal that last-token embeddings for `Act Pas` (left) and `Prep Phrase` (right) progressively merge across layers in both Gemma (top) and Llama (bottom). Light blue and dark blue dots represent the last-token embeddings for the first and second sentences in a matched pair, respectively. For Gemma, layers 1, 6, 12, and 18 are shown, and for Llama, layers 1, 11, 22, and 32 are shown. Layer numbers are displayed in the top left corner of each subplot.

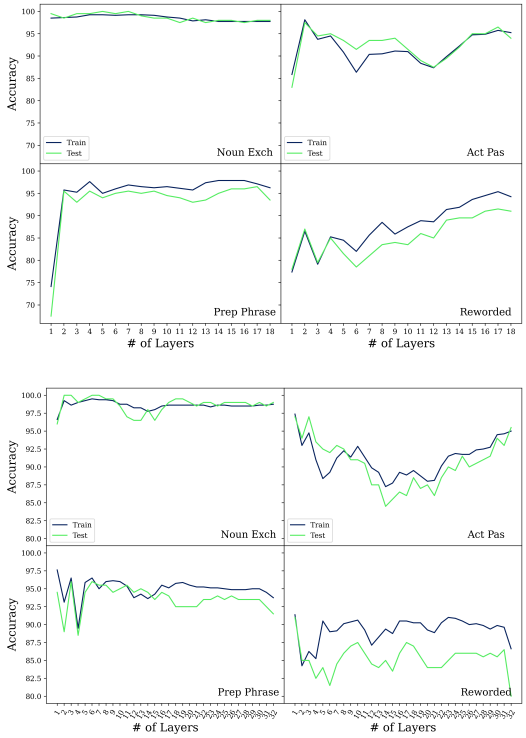

Figure 4: Train and test accuracy of the logistic regression probe across Gemma and Llama layers demonstrate the models effectively distinguish structurally similar and dissimilar sentences. Training and test accuracy are shown in dark blue and light green respectively.

purely semantic similarity, particularly for `Act Pas` and `Prep Phrase`. The PCA results likely reflect local processing in shallow layers and structural integration in deeper layers.

When projecting the difference vectors between the last token embeddings of matched and unmatched pairs, matched pairs form tight clusters, while unmatched pairs are more dispersed, across the model layers. This indicates higher similarity in semantic meaning and sentence structure for matched pairs.

Probing We recorded both training and testing accuracy for Gemma and Llama across model layers, as shown in Fig. 4. The probing accuracy for `Noun Exch` and `Prep Phrase` exceeded 90% in both models, while the accuracy for `Act Pas` and `Reworded` was slightly lower, around 80%. These results suggest that the models can effectively distinguish between structurally similar and dissimilar sentences. Notably, for `Act Pas` in Llama, matched and unmatched sentence pairs mingle in the intermediate layers, likely because the processing of semantic meaning occurs in these layers. In the deeper layers, however, the norm of differences between matched pairs approaches zero, while unmatched pairs maintain a larger norm, leading to better separation and improved probe performance.

4 Conclusion

Our findings provide insights into how LLMs process semantic meaning across layers. In both Gemma and Llama, the ability to reconcile pragmatic meaning between semantically similar sentences is influenced by both semantic complexity and sentence structure and sentences with high similarity in semantic meaning and structures consistently yield highly similar last-token embeddings. In Gemma, trends in cosine similarity, norm ratio, and PCA visualizations demonstrate the model's ability to distinguish nuanced differences while maintaining an understanding of shared meaning between sentence pairs. Llama, however, exhibits a two-stage process: earlier layers primarily encode information locally, handling contextual and structural differences, while deeper layers integrate and reconcile these distinctions, as indicated by the dip and subsequent stabilization in cosine similarity. In both models, full-sentence comprehension gradually emerges in the deeper layers.

Limitation

We acknowledge several limitations in our study. First, the simplicity and brevity of the sentences used may restrict the generalizability of our findings to more complex linguistic constructs. Additionally, using only 500 sentence pairs per type raises concerns about representativeness and potential biases in our sample, particularly for the `Reworded` sentence type. Our analysis also heavily relies on qualitative methods, particularly visualizations, to draw conclusions. While these visual interpretations provide valuable insights, they remain speculative and are influenced by pre-existing intuitions. Incorporating more robust quantitative metrics would improve the validity and reliability of our results, offering a more comprehensive understanding of model behavior.

Furthermore, the use of cosine similarity and norm ratio depends on the Euclidean inner product, which may not always be the most appropriate choice for all models. While it appears to work well for Llama, it doesn't perform as effectively in Gemma, where embeddings are already very close to each other. An alternative, such as the causal inner product introduced by Park et al. (2024b), could be applied, though it currently has the limitation of only being applicable to the last layer. Extending this causal inner product to intermediate layers could provide more concrete results and enhance the objectivity and rigor of our findings.

References

ACM FAccT. 2023. Policy on harmful research. https://facctconference.org/2023/harm-policy. Accessed: 2024-09-30.

AI@Meta. 2024. Llama 3 model card.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Gunther, Anton Korinek, José Hernández-Orallo, Lewis Hammond, Eric J. Bigelow, Alexander Pan, Lauro Langosco di Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Se'an 'O h'Eigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob N. Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *ArXiv*, abs/2404.09932.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian K. Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Gunecs Baydin, Sheila A. McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Markus Brauner, and Sören Mindermann. 2023. Managing ai risks in an era of rapid progress. *ArXiv*, abs/2310.17688.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *ArXiv*, abs/2309.12288.

CAIS. 2023. Statement on ai risk. https://www.safe.ai/work/statement-on-ai-risk. Accessed: 2024-09-30.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. 2024. Not all language model features are linear. *ArXiv*, abs/2405.14860.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Hila Gonen, Terra Blevins, Alisa Liu, Luke Zettlemoyer, and Noah A. Smith. 2024. Does Liking Yellow Imply Driving a School Bus? Semantic Leakage in Language Models. *arXiv e-prints*, arXiv:2408.06518.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *ArXiv*, abs/2310.02207.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2024. Exploring concept depth: How large language models acquire knowledge at different layers? *ArXiv*, abs/2404.07066.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them:

Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14551–14558, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of lexical stylistic features in language models' embedding space. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. *ArXiv*, abs/2309.00941.

OpenAI. 2024. Gpt-4 technical report. *ArXiv*.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024a. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024b. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*.

Tiago Pimentel, Rowan Hall Maudslay, Damian Blasi, and Ryan Cotterell. 2020. Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015, Online. Association for Computational Linguistics.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. 2024. Towards compositionality in concept learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46530–46555. PMLR.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *ArXiv*, abs/2310.15154.

Rocco Tripodi. 2021. How contextualized word embeddings represent word senses. In *Italian Conference on Computational Linguistics*.

Jacob Turton, David P. Vinson, and Robert Smith. 2020. Deriving contextualised semantic features from bert (and other transformer model) embeddings. *ArXiv*, abs/2012.15353.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Conference on Empirical Methods in Natural Language Processing*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *ArXiv*, abs/2305.14160.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Cheng Zhang, Stefan Bauer, Paul N. Bennett, Jian chuan Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding causality with large language models: Feasibility and opportunities. *ArXiv*, abs/2304.05524.

Appendix

A Related Work

The study of semantic meaning in LLMs can be divided into lexical and pragmatic semantics. While lexical semantics focuses on how individual word meanings are encoded, pragmatic semantics examines how LLMs process meaning beyond literal definitions. Much research has explored lexical semantics in BERT-like models, showing that they handle lexical ambiguity by reducing contextual uncertainty and effectively representing polysemy (Pimentel et al., 2020; Garí Soler and Apidianaki, 2021). Furthermore, Ravichander et al. (2020) and Vulić et al. (2020) explored how BERT models capture lexical knowledge, such as hypernymy, with probing tasks.

While lexical semantics has been extensively studied, pragmatic semantics remains a challenge for LLMs interpretability. Recent research suggests that LLMs encode linguistic information and broader concepts such as space, time, and world models in self-supervised sequence tasks (Gurnee and Tegmark, 2023; Nanda et al., 2023). However, LLMs often struggle with more nuanced tasks, such as bridging inferences and resolving ambiguity, which reveals limitations in their causal reasoning and contextual understanding (Zhang et al., 2023; Berglund et al., 2023; Gonen et al., 2024). Specifically, Gonen et al. (2024) highlighted the issue of semantic leakage, where LLMs infer unintended associations from contextually unsupported information. These findings emphasize the need for deeper insight into how LLMs manage pragmatic meaning and make inferences beyond what is explicitly stated.

In efforts to understand how LLMs organize semantic concepts, the linear representation hypothesis posits that high-level concepts are encoded linearly within LLMs' embedding spaces, a notion originating from early work on continuous word representations (Mikolov et al., 2013). Building on this, recent studies have shown that attributes such as sentiment and truthfulness follow linear structures within the embedding space (Tigges et al., 2023; Marks and Tegmark, 2023). Expanding on this idea, researchers have also explored more complex geometric structures in LLM representations, demonstrating that categorical concepts form simplices, hierarchical ones exhibit orthogonal structures (Park et al., 2024a), some features are encoded circularly (Engels et al., 2024), and LLMs organize complex ideas non-linearly from simpler components (Stein et al., 2024).

Beyond the geometric structure of semantic knowledge, progress has also been made in understanding how LLMs process information hierarchically through their layers. Earlier work on BERT-like models demonstrated that shallow layers primarily gather contextual information, while deeper layers focus on prediction tasks (Voita et al., 2019; Wang et al., 2023). Wang et al. (2023) also found that semantic information consolidates into label word representations in deeper layers, supporting final predictions. Similarly, Lyu et al. (2023) showed that lower layers in BERT-based models tend to encode basic lexical meanings, while higher layers capture stylistic and associative nuances. Advancing this research, Liu et al. (2024) examined generative LLMs and found that lexical semantics are primarily encoded in the earlier layers, while higher layers shift toward prediction, resulting in weaker semantic induction. This aligns with Jin et al. (2024)'s "Concept Depth" hypothesis, which proposes that abstract concepts—such as factual and inferential knowledge—emerge progressively in deeper layers.

B Dataset Details

See Table 1 for example sentences and completions by Gemma and Llama respectively. We suppose `Noun Exch` sentences share the highest semantic similarity. Though `Act Pas` sentences technically share the same semantic meaning, passive voice sentences are less commonly used and may carry slightly different stylistic nuances. Similarly, the pair in `Prep Phrase` maintain semantic equivalence but also differ in frequency of usage.

Type	Example	Appended Term	Completions
Noun Exch	Evelyn and Simon watched a movie. Simon and Evelyn watched a movie.	They	spent 1/2 of the movie watching. spent $12.50 on popcorn. saw the movie for the first time in their lives. were both very excited.
Act Pas	The scientist conducted an experiment. An experiment was conducted by the scientist.	And	the results were not what he expected. the results are shown in the table below. it turned out that the experiment failed. he used the following equation to calculate the pressure.
Prep Phrase	He saw the painting in the gallery. In the gallery, he saw the painting.	And	he saw the painting in the museum. he was so impressed that he bought it. he took it home. in the painting, he saw himself.
Reworded	I need to finish this report by tomorrow. This report must be completed by me by tomorrow.	And	I need to do it in a way that will. I have to do it in 10 minutes. I have to go to the bank to get some. I need it to be 100% correct.

Table 1: Examples for four types of sentences and their completions. The first two completions are by Gemma, and the second two are by Llama.

C PCA Visualization

This section presents additional supplementary visualizations for PCA analysis. See Section 3 for details.

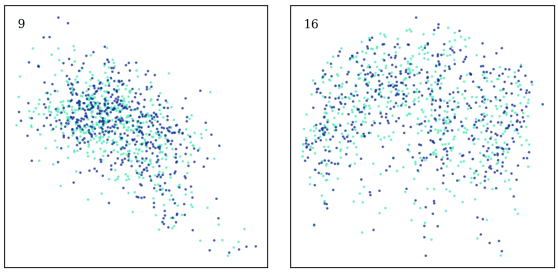

Figure 5: PCA of last token embeddings for `Noun Exch` at Gemma layer 9 (left) and Llama layer 16 (right). Light blue and dark blue dots represent the last token embeddings for the first and second sentences in a matched pair, respectively. The near-complete overlap of embeddings for the first and second sentence last token embeddings indicates the high semantic similarity between `Noun Exch` pairs.

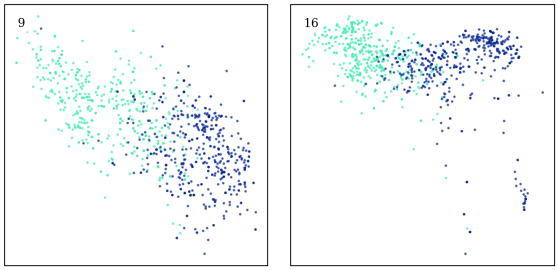

Figure 6: PCA of last token embeddings for `Reworded` at Gemma layer 9 (left) and Llama layer 16 (right). Light blue and dark blue dots represent the last token embeddings for the first and second sentences in a matched pair, respectively. Limited overlap exists between the last token embeddings of the first and second sentences in a pair. This pattern persists across model layers, suggesting that both Gemma and Llama struggle to fully reconcile the differences in meaning between `Reworded` sentence pairs, or that inherent semantic differences exist between the pairs.

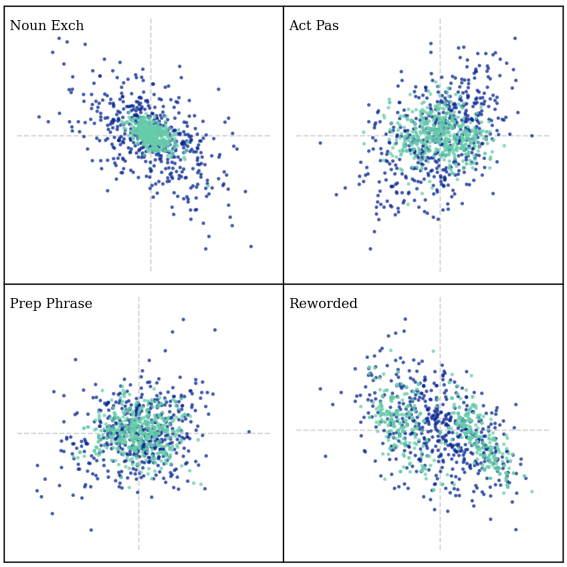

Figure 7: PCA visualization of difference vectors between matched (semantically similar) and unmatched (semantically dissimilar) sentence pairs for all four subtypes reveals no clear separation between embeddings, a pattern consistent across layers in both Gemma and Llama. The visualization for Gemma at layer 9 is shown, with Llama displaying a similar trend. Light blue dots represent difference vectors for matched pairs, while dark blue dots represent unmatched pairs. Gray dashed lines indicate the x and y axes at zero. The horizontal axis represents the average difference vector, and the vertical axis represents the average sum vector.