Modeling Ambiguity: Representation, Disambiguation, and Interpretation of Lexical Semantics in (Large) Language Models

Zhu Liu

Tsinghua University
liuzhu22@mails.tsinghua.edu.cn

Lexical ambiguity—manifested as polysemy, homonymy, and multifunctionality—is a universal linguistic phenomenon that humans effortlessly resolve through contextual cues. This review investigates how language models (LMs), as mathematical, brain-inspired systems, represent and comprehend such ambiguity, given their remarkable performance in language tasks. We synthesize existing research from two complementary perspectives: externally, by evaluating LMs on word-sense disambiguation and related tasks; and internally, by examining the interpretability of their hidden representations across different model architectures. The synthesis concludes that while language models are generally effective at capturing lexical semantics, decoder-only large language models (LLMs) exhibit unique characteristics and limitations, offering a consolidated overview for future research.

1. Introduction

A linguistic form—ranging from a morpheme or word to a phrase, sentence, or even an entire discourse—can be associated with multiple meanings. The specific meaning is modulated by its surrounding context, whether linguistic or extra-linguistic. This phenomenon, known as *ambiguity*, is universal across languages and cultures. When provided with adequate contextual cues, individuals from diverse speech communities can disambiguate these forms with remarkable ease, enabling successful communication.

As fundamental units of language, words exhibit ambiguity to varying degrees and with different levels of discriminability. This ambiguity can be systematically classified into *multifunctionality* (Haspelmath 2003), *polysemy*, and *homonymy*—an order that reflects a progressive decrease in semantic relatedness. Representative examples of these three categories are illustrated in Table 1. These distinctions are conventionally reflected in dictionary organization: multifunctionality is often either omitted or briefly indicated with a functional label (e.g., "contrasting conjunction"); polysemy is treated within a single entry but listed under different numbered senses; whereas homonymy is assigned separate entries that share the same word form.[1]

Language models mathematically formalize the mapping of linguistic tokens into continuous vector representations. Through training on vast datasets across layers containing billions of parameters, these models–particularly modern large language models (LLMs) (OpenAI 2023)–excel at text understanding and generation tasks, in-

[1] This is exemplified in Chinese lexicographical practice, where a superscript numeral is used to distinguish homonymous characters, such as "花$_1$ (flower)" and "花$_2$ (spend)".

Aspects	Level	Related	Sense items	Classical Unit	Example
Homonymy	Word	✘	✔	content words	bank, bat
Polysemy	Sense	✔	✔	content words	face, sharp
Multifunctionality	Usage	✔	✘	func., aff., partial adv.	and, again

Table 1: Taxonomy and characteristics of lexical ambiguity types. The characteristics includes the level of meaning distinction, degree of semantic relatedness, treatment as discrete sense items in lexicography, typical linguistic units involved, and representative examples. Abbreviations: "func.," "aff.," and "adv." denote function words, affixes, and adverbs, respectively. Symbols ✔ and ✘ indicate relative strength of a feature (more/less) rather than a binary presence/absence.

cluding dialogue generation (Heck et al. 2023), named entity recognition (Wang et al. 2025), and machine translation (Zhu et al. 2024). A central research question thus arises: how do these models represent and resolve *lexical ambiguity*? While numerous studies have addressed specific tasks involving contextual lexical semantics, such as word sense disambiguation (Navigli 2009), semantic similarity (Pilehvar and Camacho-Collados 2019), and co-reference resolution (Liu et al. 2023), many are limited to traditional architectures (Kenton and Toutanova 2019) or reduce disambiguation to a classification problem, thereby overlooking the inherent uncertainty and structural relationships between meanings.

This doctoral research presents a systematic investigation into the representation of lexical ambiguity in language models, structured around three interconnected strands. The first series of work advances beyond simplistic classification by employing *uncertainty estimation*, demonstrating its necessity in capturing semantic vagueness, contextual underspecification, and distributional shift. The second strand targets the complex *multifunctionality* of function words—often more semantically intricate than content words—where we develop a novel top-down method for constructing meaning-function graphs, with plans for extension to language model analysis. The third work probes the internal *mechanisms* of decoder-only LLMs, comparing them against varied architectures to identify where and how contextual meanings are captured within their hierarchical structures, thereby enhancing the interpretability of these notoriously opaque systems.

This report provides a concise overview of the aforementioned research strands, presenting key findings to summarize the core contributions of my doctoral research. The discussion concludes by listing current limitations and outlining relevant promising directions for future work. Due to space constraints, comprehensive related work and extensive experimental details have been omitted; interested readers are referred to the corresponding publications for complete technical expositions.

2. Representing Ambiguity as Uncertainty Estimation

Lexical ambiguity, where a single word possesses multiple meanings, underlies numerous cross-linguistic phenomena (see Table 1). This ambiguity can persist even within a specific context, manifesting as vagueness or enabling puns. Such cases often result in graded sense interpretations, annotator disagreement (Schlechtweg et al. 2025), and significant uncertainty when determining a word's contextual meaning. Nevertheless, traditional lexical-semantic tasks—such as Word Sense Disambiguation (WSD)[2] and Words-in-Context (WiC)[3]—are typically formalized as deterministic classification problems. This formulation presupposes a single correct answer for each word, thereby overlooking the inherent uncertainty and disagreement present in authentic language use.

To address this limitation, our prior work (Liu and Liu 2023) reframes the sense selection in Word Sense Disambiguation (WSD) as an *uncertainty estimation* (UE) problem: instead of seeking a single correct sense, we ask *how uncertain a model is when choosing among potential meanings*. We distinguish two primary sources of uncertainty:

- **Model Uncertainty** arises from limitations in the model itself, particularly when facing out-of-distribution (OOD) test data. This type of uncertainty can typically be reduced by acquiring more training data or enhancing the model's knowledge.
- **Data Uncertainty** stems from the inherent noise and ambiguity present in the data itself. This uncertainty is irreducible, persisting even with unlimited or perfectly representative data.

To systematically evaluate uncertainty, we designed controlled scenarios simulating varying degrees of data ambiguity by manipulating the contextual information around target words—defined through either syntactic dependencies or linear order. Model uncertainty was separately assessed using an out-of-distribution (OOD) dataset. We applied multiple uncertainty estimation (UE) metrics to these scenarios using a state-of-the-art model.

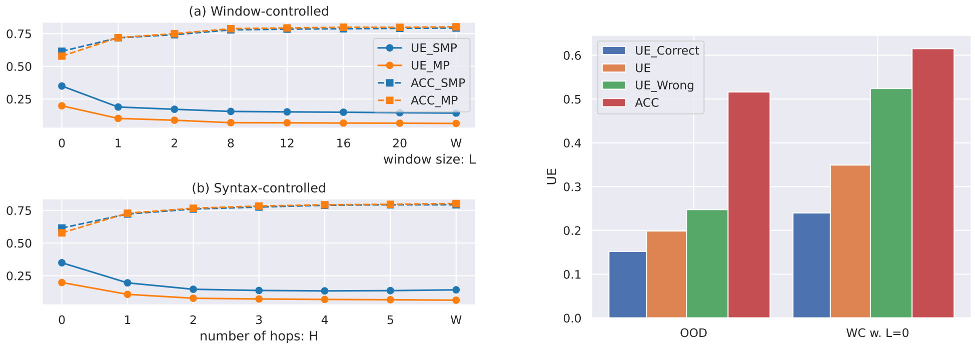
Our analysis employed the Sampled Maximum Probability (SMP) score, evaluated against two performance metrics. As illustrated in Figure 1, data uncertainty decreases incrementally as more context is provided, while model uncertainty remains consistently lower than even the minimal data uncertainty. This indicates that SMP effectively captures data uncertainty but tends to underestimate model uncertainty.

Further investigating the lexical properties influencing data uncertainty—including syntactic category, morphology, sense granularity, and semantic relations—we found that all examined properties, except for the number of synonyms, significantly impact uncertainty levels.

A related variation of WSD is Words-in-Context (WiC) (Pilehvar and Camacho-Collados 2019), which frames meaning comparison as a binary classification task. The CoMeDi shared task (Schlechtweg et al. 2025) extends this into a fine-grained four-level classification, ranging from unrelated to identical meanings. Annotator ratings are aggregated into an average score (accuracy) and standard deviation (disagreement), with systems required to predict both values for unseen instances.

In our participation (Liu, Hu, and Liu 2025), we unified these subtasks by modeling them as estimating the parameters (μ and σ) of a Gaussian distribution. The σ parameter directly captures the uncertainty in sense comparison, enabling application of our UE techniques. This approach achieved competitive performance among all submitted systems.

[2] The task of identifying the correct sense of a word in context.
[3] The task of determining if a target word shares the same meaning across two sentences.



(a) UE scores and acc. with controlled context

(b) Model vs. data uncertainty and accuracy

Figure 1: Analysis of uncertainty estimation in lexical disambiguation. (a) SMP and MP scores with F1 accuracy under context control (“0”: target word only; “W”: full context). (b) Uncertainty and accuracy comparing model uncertainty (OOD) and data uncertainty (window-controlled, *L*=0), analyzed by classification correctness.

3. Evaluating Meaning Structure Using Semantic Map Models

Semantic Map Models (SMMs) represent meanings or functions as nodes in a network, with edges indicating their associations. This approach is particularly valuable for modeling the flexible usage patterns of function words across languages. SMMs adhere to two core principles: the *Connectivity Hypothesis* (Haspelmath 2003), requiring that meanings shared by a single word form constitute a connected subgraph; and the *Economy Principle*, which minimizes redundant edges.

Traditional SMM construction follows a bottom-up approach, iteratively adjusting connections to satisfy these principles—a process that is labor-intensive and difficult to scale. In contrast, our work (Liu et al. 2025b) introduces a novel top-down graph-based algorithm. We reformulate the principles into three global constraints: (1) overall connectivity, (2) acyclicity, and (3) maximum co-occurrence weight. This reformulation transforms the problem into finding a maximum spanning tree in a fully connected graph, solvable with established algorithms like Prim’s (Prim 1957) or Kruskal’s (Kruskal 1956). We further propose a topological metric—the standard deviation of node degrees—to select the optimal tree from candidate solutions.

We validate our approach through a case study on repetitive and supplement adverbs (Guo 2010), covering 28 forms across 9 languages and 18 functions. The generated semantic map is evaluated both intrinsically—reporting graph statistics including connectivity hypothesis satisfaction—and extrinsically against linguist-annotated ground truth. As shown in Table 2, our algorithm demonstrates strong effectiveness and efficiency.

To support practical application, we have developed an interactive visualization tool (Liu et al. 2025a) to assist typologists in data analysis and presentation, with positive feedback from user surveys.

As an ongoing extension, we are integrating language models to represent forms across functions, aiming to enhance initial graph construction efficiency. Conversely,

Model	Size↑	Recall↑	Precision↑	Accuracy↑
C	286	1.00	0.00	50.0
LT	-	-	-	79.0
GT	91	1.00	0.20	1.00
0	90	85.7	0.17	**92.6**
1	89	82.1	0.21	91.4
2	89	82.1	0.44	90.1
3	88	82.1	0.34	91.4
4	88	78.6	0.50	88.9

Table 2: Performance comparison of generated semantic maps against baselines: complete graph (C), ground truth (GT), and literature standard (LT). Index 0-4 denotes top candidate maximum spanning trees (×10,000). Accuracy measures alignment with GT.

the typologist-curated meaning networks provide a challenging structured evaluation framework for assessing language model capabilities.

Despite the impressive performance of large language models (LLMs), their internal mechanisms for lexical semantic processing remain opaque. Unlike earlier specialized models (Word2Vec, GloVe) or encoder-only architectures like BERT, decoder-only LLMs employ a unified next-token prediction objective that intertwines understanding and generation, complicating interpretation.

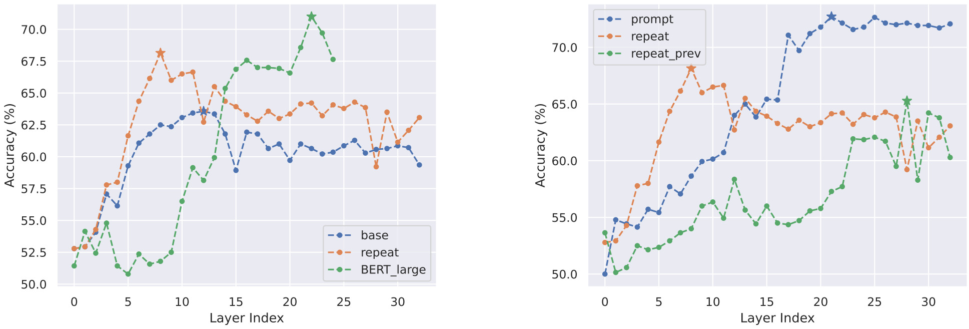
In our ACL 2024 work (Liu et al. 2024), we analyze layer-wise representations of decoder-only LLMs versus encoder-only BERT on the WiC task. As shown in Figure 2a, decoder models exhibit an inverted U-shaped performance curve, peaking in early-to-middle layers, while encoder models show monotonically improving performance toward higher layers. This pattern suggests decoder LLMs establish word understanding in lower layers and shift to prediction tasks in higher layers, aligning with their training objective.

In ongoing work (Liu et al. 2025c), we investigate the relational structure of token embeddings in LLMs by constructing connectivity-constrained networks across the vocabulary. Preliminary analysis reveals strong small-world effects—characterized by short path lengths between tokens—with larger models exhibiting more pronounced effects than smaller counterparts. This work aims to extend the analysis to additional models to further validate these observations.

4. Conclusion and Limitations

This report has summarized my doctoral research on how language models—particularly decoder-only LLMs—represent, disambiguate, and learn lexically ambiguous meanings. Through reformulations of classical tasks, typologically-inspired network analysis, and representation probing, we demonstrate that language models effectively capture lexical semantics, while highlighting the unique characteristics of decoder-only architectures.

Several limitations warrant further investigation. First, despite some multilingual evaluation, our focus remains predominantly on English, potentially overlooking language-specific ambiguity patterns. For instance, Chinese exhibits distinct morphological structures from English, which directly influences its lexical semantic properties. Second, our analysis operates primarily at the representation level, neglecting finer-grained mechanistic interpretations at the neuronal level (Olah 2022). Third, while



(a) Layer-wise accuracy across models

(b) Layer-wise accuracy of Llama2

Figure 2: Layer-wise representation analysis. Representation extraction methods: base/repeat (target word), prompting (last punctuation), repeat_prev (previous word).

probing-based evaluation is widely adopted, its reliance on task-specific performance raises questions about generalizability and predictive power. Future work will address these constraints by expanding linguistic coverage, incorporating mechanistic analysis, and exploring more integrated semantic frameworks such as functional distributional semantics (Emerson and Copestake 2016).

References

Emerson, Guy and Ann Copestake. 2016. Functional distributional semantics. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 40–52, Association for Computational Linguistics, Berlin, Germany.

Guo, Rui. 2010. A semantic map study of adverbs related to "supplement". In *Paper presented at the International Symposium for Comparative and Typological Research on Languages of China*, Hong Kong, China.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The new psychology of language*, volume 2. Lawrence Erlbaum, Mahwah, NJ, pages 211–243.

Heck, Michael, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Association for Computational Linguistics, Toronto, Canada.

Kenton, Jacob Devlin Ming-Wei Chang and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kruskal, Joseph B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50.

Liu, Ruicheng, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Liu, Zhu, Zhen Hu, Lei Dai, and Ying Liu. 2025a. Xism: an exploratory and interactive graph tool to visualize and evaluate semantic map models. https://arxiv.org/abs/2507.04070.

Liu, Zhu, Zhen Hu, and Ying Liu. 2025. JuniperLiu at CoMeDi shared task: Models as annotators in lexical semantics disagreements. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 103–112, International Committee on Computational Linguistics, Abu Dhabi, UAE.

Liu, Zhu, 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, Association for Computational Linguistics, Bangkok, Thailand.

Liu, Zhu, Cunliang Kong, Ying Liu, and Maosong Sun. 2025b. A top-down graph-based tool for modeling classical semantic maps: A case study of supplementary adverbs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4567–4576, Association for Computational Linguistics, Albuquerque, New Mexico.

Liu, Zhu and Ying Liu. 2023. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3963–3977, Association for Computational Linguistics, Toronto, Canada.

Liu, Zhu, Ying Liu, KangYang Luo, Cunliang Kong, and Maosong Sun. 2025c. From the new world of word embeddings: A comparative study of small-world lexico-semantic networks in llms. https://arxiv.org/abs/2502.11380.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Olah, Chris. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2(4).

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

Prim, Robert C. 1957. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Schlechtweg, Dominik, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, International Committee on Computational Linguistics, Abu Dhabi, UAE.

Wang, Shuhe, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Association for Computational Linguistics, Albuquerque, New Mexico.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Association for Computational Linguistics, Mexico City, Mexico.