

Trick or Neat: Adversarial Ambiguity and Language Model Evaluation

Antonia Karamolegkou¹ Oliver Eberle²

Phillip Rust¹ Carina Kauf³ ✉ Anders Søgaard¹

¹University of Copenhagen ²Technische Universität Berlin

³Massachusetts Institute of Technology ✉ Aleph Alpha Research

Correspondence: antka@di.ku.dk

Abstract

Detecting ambiguity is important for language understanding, including uncertainty estimation, humour detection, and processing garden path sentences. We assess language models' sensitivity to ambiguity by introducing an adversarial ambiguity dataset that includes syntactic, lexical, and phonological ambiguities along with adversarial variations (e.g., word-order changes, synonym replacements, and random-based alterations). Our findings show that direct prompting fails to robustly identify ambiguity, while linear probes trained on model representations can decode ambiguity with high accuracy, sometimes exceeding 90%. Our results offer insights into the prompting paradigm and how language models encode ambiguity at different layers. We release both our code and data: [coastalcph/lm_ambiguity](https://github.com/coastalcph/lm_ambiguity).

1 Introduction

Linguistic utterances often have ambiguous meanings, but our world knowledge helps us resolve them. Consider the ambiguous sentence in (1), and the unambiguous alterations in (2) and (3):

- (1) The man saw the woman with the telescope.
- (2) The man saw the woman with the dress.
- (3) The man saw the woman with his own eyes.

Even though sentences of the form *NP V'ed NP with NP* are always *structurally* ambiguous between an interpretation where the PP modifies the object NP – the so-called *low attachment* reading exemplified by (2) – and a reading where the PP is the instrument of the VP – the *high attachment* reading exemplified by (3) – our world knowledge can help us resolve such ambiguities by ruling out unlikely interpretations. In particular, although it is arguably roughly equally conceivable to do both interpretations in (1) rendering both structurally and semantically ambiguous, in the absence of licensing context, the sentences in (2) and (3) are

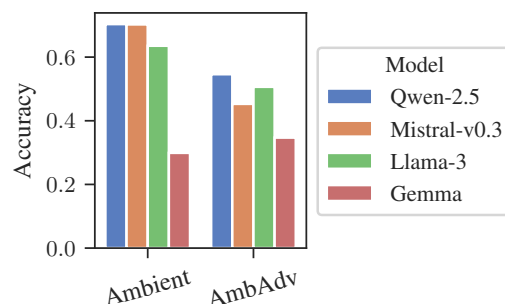


Figure 1: Results across models for an existing ambiguity dataset, Ambient, and our adversarial ambiguity dataset, AmbAdv.

semantically unambiguous, because it is implausible for someone to use a dress as a seeing device, and implausible for someone to walk around with another person's eyes. Once supporting information is introduced, however, implausible sentences can become plausible (Nieuwland and Van Berkum, 2006). For example, a context in which someone calls a telescope a dress or mistakes a telescope for a dress or has a dress-shaped telescope can make the instrument-PP interpretation in (2) more likely, revealing the sentence's underlying structural ambiguity. In the absence of such licensing context, reading time studies show that humans have a general preference to attach an ambiguous PP to the VP rather than the NP (high attachment over low attachment), with higher reading times around the PP region, when the high attachment reading is *not* semantically licensed (Spivey-Knowlton and Sedivy, 1995), such as in (2).

Ambiguity in language poses a great challenge for Language Models (LMs) (see Figure 1; Liu et al., 2023), as it can lead to hallucinations, biased completions, and misinterpretations. This can degrade performance on tasks such as fact-checking, sentiment analysis, and information extraction (Keluskar et al., 2024). Different types of

ambiguity affect models in distinct ways: syntactic ambiguity (e.g., (1)) can lead to parsing errors and flawed summarization or translation; lexical ambiguity (e.g., “the speaker”) can confuse question answering and retrieval systems, especially in low-context scenarios; and phonological ambiguity (e.g., “ice cream” vs. “I scream”) complicates speech recognition and dialogue modeling. Models that can detect and reason about ambiguity can help mitigate these issues by flagging unclear inputs, suggesting alternative interpretations, and improving user understanding—particularly in sensitive contexts like misleading news headlines or political claims (Liu et al., 2023). Ambiguity-aware modeling is therefore essential for generating coherent, trustworthy outputs that align with human expectations (Kamath et al., 2024a).

To systematically evaluate how well LMs handle ambiguity, we introduce AmbAdv, a dataset with adversarial variations of ambiguous sentences, and find that most LLMs struggle to identify ambiguity, with some exhibiting a ‘yes bias’ when prompted, highlighting the need to account for class distribution in evaluation. We examine how LLMs encode ambiguous sentences and find that linear probes on layer-wise representations can reliably distinguish ambiguity. Our analysis of model disambiguations and representations suggests that LLMs may partially rely on memorization to resolve ambiguity.

2 Related Work

Ambiguity enables efficient communication by relying on context and minimizing processing load (Piantadosi et al., 2011). Previous works have already highlighted ambiguity as a challenge in a variety of NLP tasks such as multimodal machine translation (Li et al., 2022), visual question answering (Stengel-Eskin et al., 2023), misleading claim detection (Liu et al., 2023) speech-to-text transcription (Zhu et al., 2024), humour style classification (Kenneth et al., 2024), sentiment analysis (Buscemi and Proverbio, 2024), semantic parsing (Stengel-Eskin et al., 2024), etc. There exist a few attempts to create ambiguity-inclusive datasets. Min et al. (2020) introduce AMBIGQA, identifying all possible answers to questions and rephrasing/disambiguating them. Liu et al. (2023) create AMBIENT, an NLI benchmark for ambiguity detection and disambiguation. Kamath et al. (2024b) create a dataset of 1,000 scope-ambiguous sentences, showing that models may be sensitive

to the meaning ambiguity.

Building on these challenges, we construct an adversarial ambiguity dataset to evaluate whether models can detect ambiguity in inputs resembling real-world user queries. While most prior datasets focus on lexical or syntactic ambiguity, phonological ambiguity remains underexplored. Recent speech modeling studies show how transformer-based models like Whisper and Wav2Vec2 handle phonological variation, including homophones (Mohebbi et al., 2023), phonotactic patterns (de Heer Kloots and Zuidema, 2024), and assimilation (Pouw et al., 2024). These findings suggest that neural models can encode ambiguity-relevant distinctions in the speech domain, motivating our inclusion of phonological ambiguity and rhyme-based perturbations. We position AmbAdv relative to existing datasets in Table 4 and provide further motivation in Appendix A.

3 Methodology

Dataset creation We construct a dataset based on 8 syntactically, 16 lexically, and 16 phonologically ambiguous sentences. These sentences were hand-picked based on existing textbooks, online linguistic studies, and ambiguity datasets (Taha, 1983; Liu et al., 2023; Stengel-Eskin et al., 2024).

Modification Type	Syntactic	Lexical	Phonological
Original	8	16	16
Word Order	8	16	-
Synonym	540	64	64
Random	404	64	64
Rhyme	137	64	64
Total	1097	224	208

Table 1: Count of adversarial modifications across different linguistic ambiguity types.

We modified the originally ambiguous base sentences to create sentence variants each using four different manipulation types: (i) *Word order*. We swap the subject and the object of the sentence or, when not possible, create passive/active voice alternations. Critically, this manipulation does not lead to a semantically implausible interpretation of either of the PP attachment possibilities. (ii) *Synonymous word*. We exchange a key word in the sentence for a synonym. Because we use synonyms, these sentence variants have the same ambiguity structure as the original example. (iii) *Random word*. We exchange a keyword in the sentence for a random word of the same syntactic category. These

sentence variants rule out one interpretation of the sentence (i.e., here: the PP can no longer be interpreted as introducing the instrument of the seeing action). (iv) *Rhyme word*. We exchange a key word in the sentence for a word that rhymes in sound with the original word.

Table 2 shows an example of these data manipulations, and Table 1 presents the dataset statistics. The total number of sentences in AmbAdv is 1529, with 671 labeled as ambiguous. Further details can be found in Appendix B. Each manipulation type targets a different type of ambiguity and model robustness. Word substitutions have long been used in adversarial NLP (Zhou et al., 2021; Goyal et al., 2023; Bespalov et al., 2023; Sabir et al., 2023; Zhang et al., 2024) to explore model vulnerabilities and assess generalization under distribution shifts. Word order can affect model performance (Abdou et al., 2022). Synonym substitutions simulate natural lexical variation while preserving ambiguity, testing semantic resilience (Hsieh et al., 2019). Random substitutions disrupt semantic coherence to test model reliance on lexical cues. Rhyme-based substitutions, introduced here as a novel perturbation, explore phonological similarity, an underexplored adversarial strategy in NLP (Suvama et al., 2024). This is particularly relevant given the growing interest in phonological effects in LLM outputs, especially in multimodal contexts (Fathullah et al., 2024).

Manipulation	Example sentence	Ambiguous?	
		Struct.	Semant.
Original	The man saw the woman with the telescope.	✓	✓
Word Order	The woman saw the man with the telescope.	✓	✓
Synonym Word	The man saw the woman with the <i>binoculars</i> .	✓	✓
Random Word	The man saw the woman with the <i>book</i> .	✓	✗
Rhyme Word	The man saw the woman with the <i>gyroscope</i> .	✓	✗

Table 2: Overview of manipulation types for the syntactically ambiguous sentences.

Dataset validation Lexical and phonological ambiguities were straightforward to annotate and validate. Lexical ambiguity was verified via dictionary lookup, while phonological ambiguity was confirmed using IPA transcriptions from authoritative sources to identify homophones or rhyming patterns. On the contrary, the syntactic sentences are all syntactically ambiguous, allowing multiple interpretations of their structure. However, considering the meanings of the words can often disambiguate them. For example, “*The man saw the woman with the dress*” (a random word substi-

tution) is structurally ambiguous, but our world knowledge precludes interpreting a dress as a visual instrument.¹ Based on this extra-linguistic reasoning, we classified sentences with random and rhyme word substitutions as *unambiguous*.

To validate our annotations, three annotators independently labeled a random 50% sample of the dataset as ambiguous or unambiguous, incorporating real-world knowledge in their judgments. Inter-annotator agreement, measured by Cohen’s Kappa, was 91%. Agreement for synonym-based sentences was ~84%, and for random/rhyme substitutions, around ~98%.

Models We are interested in evaluating whether LLMs can reliably modulate their judgments of a sentence’s ambiguity status in the face of minimal adversarial attacks that either change or do not change the sentence’s ground truth ambiguity label. To this end, we use four open-access instruction-tuned LLMs in the 7 Billion parameter regime, selected based on their performance on the LMSys chatbot arena leaderboard:² Qwen-2.5-7b, Mistral-7b-v0.3, Llama-3-7b, and Gemma-7b. We set the experiments in a binary classification set-up because it provides a straightforward and interpretable approach, which can also be seen in a real-world scenario in which a user asks a model if a sentence (e.g., a news headline) is ambiguous.

Prompting strategy We frame our experiments as ambiguity identification tasks. We use 8 different prompts that elicit responses in a structured Jinja2 format across the different model chat templates.³ The prompts are: 2 default prompts and 6 prompts with various binary words in different orders to investigate a potential “yes-bias” (Dentella et al., 2023) and model consistency.⁴ For a set of 120 sentences, we also use 2 disambiguation prompts. We provide the templates in Appendix C.1. To evaluate the model responses, we calculate the accuracy/error rate, i.e., the proportion of matches/mismatches between the annotated and predicted ambiguity status.

¹Except, of course, if the dress has a mirror. Such exceptions illustrate how pragmatics may override semantics (Morris, 1946), but LLMs must be sensitive to distinctions between syntactic ambiguities that support multiple conventional readings and those typically resolved by context.

²  [spaces/lmsys/chatbot-arena-leaderboard](https://github.com/lmsys/chatbot-arena-leaderboard)

³  [jndiogo/LLM-chat-templates](https://github.com/jndiogo/LLM-chat-templates)

⁴We define consistency here as the ability to make reliable decisions in semantically similar contexts demonstrating a systematic capacity to generalize across language variations” (Elazar et al., 2021).

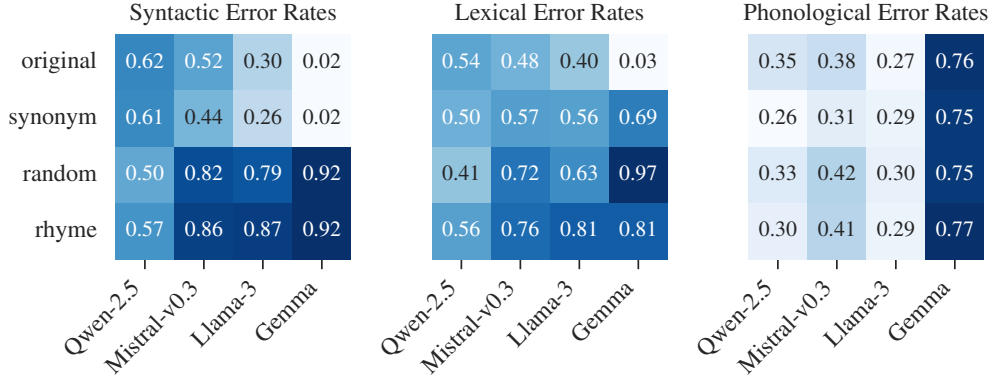


Figure 2: Results across models after zero-shot prompting, showing the error rate of the different synonym, random, and rhyme word modifications. Values closer to 1 indicate a higher error rate.

	Syntactic	Lexical	Phonetic	Ambient
Qwen-2.5	0.436 \pm 0.028	0.491 \pm 0.128	0.781 \pm 0.032	0.693 \pm 0.025
Mistral-v0.3	0.360 \pm 0.064	0.407 \pm 0.088	0.647 \pm 0.197	0.703 \pm 0.055
Llama-3	0.462 \pm 0.026	0.360 \pm 0.058	0.697 \pm 0.105	0.636 \pm 0.068
Gemma	0.525 \pm 0.015	0.268 \pm 0.006	0.247 \pm 0.092	0.299 \pm 0.080

Table 3: Average accuracy results across 8 different prompt templates.

4 Experiments and Results

Ambiguity Identification We visualize the average results over five runs across 8 prompts in Figure 1 and Table 3. We observe several trends: (1) Most models perform worse on AmbAdv, even though the sentences are simpler compared to Ambient, indicating some model sensitivity in the adversarial sentence modifications. (2) The models demonstrate consistent performance across 8 different prompts, with an average accuracy difference of less than 0.1 points. (3) Qwen and Mistral perform worst in the syntactically ambiguous sentences, while Llama performs worst in the lexical ambiguous sentences. (4) Gemma performs better on the syntactic set, but upon checking the distribution of responses, we found that it gives almost all affirmative answers. The syntactic set is more balanced with equal positive and negative examples, while the other datasets contain more negative examples (see Figure 4). (5) This suggests a “yes bias” in most of the model outputs.

Adversarial Sentence modifications. We provide an aggregation of each model’s performance across types of sentence modifications in Figure 2. We observe the following: (1) Ambiguity identification seems to be a challenging task for most models. (2) Synonym words are not always more challenging than the original sentences. (3) Both random

and rhyming words pose the greatest challenge for most models, suggesting a lack of resilience to adversarial substitutions. (4) Qwen is the only model that performs better on random substitutions and struggles with the original sentences. This behavior may suggest an attempt to avoid memorization of the original sentences, which were likely part of its training data. Alternatively, it could indicate that the model maintains a more balanced approach in generating yes/no responses.

Disambiguation analysis For the syntactic ambiguity set, we prompted models to also provide disambiguations on a set of 120 sentences. Manual error analysis revealed the following: (1) Only Qwen and Llama provided accurate disambiguations, but only for the original and synonym cases. Mistral and Gemma add hallucinations in their responses (e.g., ‘The man saw a woman near the location of the dress’, or ‘The man saw the woman with the camera, and she was taking pictures’). (2) Despite identifying ambiguity, many models lack world knowledge—e.g., they change the action verb with synonym instruments for ‘telescope’ (e.g., ‘using or holding a camera/glasses/polaroid’ but not ‘seeing with a camera/glasses/polaroid’). Additionally, inanimate objects are sometimes assigned actions (e.g., ‘A mug witnessed a woman using a telescope’, ‘The telescope saw the man who was with the woman’). (3) In some cases,

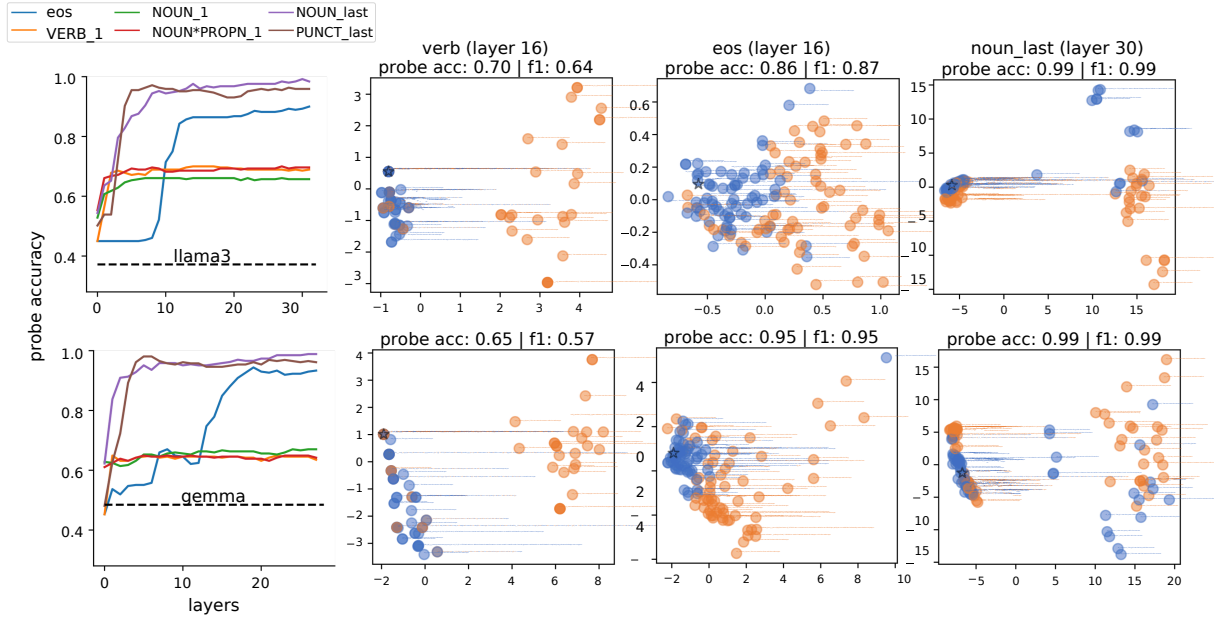


Figure 3: Representational analysis of hidden state representation across layers. Linear Probe accuracies for Llama-3 and Gemma for different functional roles of tokens, e.g., first verb or last noun (left). PCA projections extracted for different token roles, blue/orange indicate ambiguous/unambiguous sentences (right).

the model repeated the disambiguation of the original sentence in synonym substitutions, indicating a memorization effect. We provide a sample of the model responses in the Appendix Tables 9 and 10.

Representational Analysis After assessing the behavioural model performance and error patterns, we now investigate to what extent residual stream representations can decode ambiguity in a subset of 142 samples of the AmbAdv dataset that contains variations of a sentence (1). We extract layer-wise representations of specific token roles, i.e., first verb, first noun/proper noun, last noun, last punctuation, and the final end of prompt sentence representation (eos). We compute probe accuracies (Radford et al., 2021; Campbell et al., 2023) for each role via a 5-fold linear probe evaluation using a logistic regression model (80/20 train/test splits). We present a summary of our results in Figure 3, with additional details provided in Appendix D. Our findings show that averaged probe accuracies are consistently higher compared to the performance of the model’s generated responses through prompting. We find probe accuracies of 0.9 and higher for token representations related to the last punctuation and last noun token, starting from layer 5 on. The first noun/proper noun and verb reach moderately high accuracy scores of 0.6 – 0.7. The final prompt token, used by the language modeling head to generate the answer

token, achieves probe accuracies of 0.85 – 0.90 from layer 12 on, which appear delayed compared to the other token types considered. An analysis of PCA projections of representations highlights that ambiguous/unambiguous (blue/orange in Figure 3) samples appear clustered, yet we do not find a clear direction that would uniquely code for ambiguity. Interestingly, we observe groups of both ambiguous and unambiguous sentences clustering around the original sentence (indicated by a blue star). This offers representational insight into repeated misclassifications, which may stem from memorized patterns, as many AmbAdv source sentences are publicly available online.

5 Conclusion

We introduce the first adversarial dataset for linguistic ambiguity to evaluate whether LLMs can assess a sentence’s ambiguity status under minimal adversarial attacks that may or may not alter its ground truth label. Our results show that LLMs struggle to accurately interpret ambiguous sentences. While ambiguity-related information seems present in the models’ representations, they fail to leverage it effectively in their outputs. Our representational analysis shows that linear probes on layer-wise representations can reliably distinguish ambiguity, suggesting that LLMs may partially rely on memorization rather than ambiguity understanding.


6 Limitations

We use prompting as our evaluation paradigm to explore the model and prompting limitations, as this method is likely to be used by users of language models in real-world settings. For example, a user might ask if a given sentence (e.g., a news headline) is ambiguous. This is why we designed the dataset with different types of linguistic ambiguity sentences, aiming to systematically investigate how well models can identify ambiguity even in publicly available and adversarial examples. Before the LLM era, linguistic knowledge encoded in neural language models and LLMs has been evaluated using either log likelihood comparisons of minimal pair sentences (Linzen et al., 2016; Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020, 2024) or through probing of the model’s representations of a stimulus (Hewitt and Manning, 2019; Eisape et al., 2022; Müller-Eberstein et al., 2022). Similarly, semantic plausibility has been evaluated using log-likelihood measures and representation probing (Kauf et al., 2023; Michaelov et al., 2023; Misra et al., 2024).

More recently, *prompting* emerged as a way to directly prompt LLMs for linguistic knowledge using natural language (e.g., Brown et al., 2020; Blevins et al., 2023). Nevertheless, a direct comparison of log-likelihood and prompting measures shows that prompting may systematically underestimate the model’s true linguistic capabilities because it requires the models not only to solve the task, but also to correctly interpret the prompt and to translate their answer into the desired output format (Hu and Levy, 2023; Hu et al., 2024). This is the reason that motivated us to include a representational analysis, in addition to a prompting-type analysis.

Moreover, adversarial datasets, having been specifically designed to fool a model, may be subject to the biases of the dataset creators (Li and Michael, 2022). However, the purpose of our dataset is not to train and build models that are more robust to spurious correlations but rather to interpret and evaluate certain model behaviours. The size of our dataset may also be a limitation, as it can only be used for cases that evaluate the sensitivity of LMs toward ambiguity and adversarial examples. Lastly, a major limitation is that our dataset only includes English sentences, which limits its applicability to other languages. We provide further motivation for our study and dataset choices in Appendix A.

7 Ethics

We do not foresee any ethical concerns. On the contrary, the scope of our dataset is purely for scientific research of language models and can potentially help identify ambiguity in political claims, news headlines, and other domains. The research was conducted in accordance with ethical principles, and no sensitive or personal data was used or collected during the study. We fairly compensated each annotator involved in this study at a rate of \$18 per hour on the crowdsourcing platform Prolific.⁵ The only requirement for annotators’ demographic and geographic characteristics was being a native English speaker. The instructions given to the annotators were very similar to the default prompts we used in our work (see Appendix C.1. The dataset contains linguistic sentences that can be found in grammar books and do not raise any privacy or ethical concerns. In terms of resources we estimate less than 12 hours GPU (A100 40GB) usage. Both our dataset and code are publicly available under CC BY 4.0⁶ and MIT licences respectively at  [coastalcph/lm_ambiguity](https://github.com/coastalcph/lm_ambiguity).

References

- Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. [Word order does matter and shuffled language models know it](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.
- Dmitriy Beshpalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou, and Yanjun Qi. 2023. [Towards building a robust toxicity predictor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 581–598, Toronto, Canada. Association for Computational Linguistics.
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. [Prompting language models for linguistic structure](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

⁵<https://www.prolific.com/>

⁶<https://creativecommons.org/licenses/by/4.0/>

- Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Alessio Buscemi and Daniele Proverbio. 2024. [Chatgpt vs gemini vs llama on multilingual sentiment analysis](#). *arXiv preprint*.
- James Campbell, Phillip Guo, and Richard Ren. 2023. [Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching](#). In *Socially Responsible Language Modelling Research*.
- Marianne de Heer Kloots and Willem Zuidema. 2024. [Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2.0](#). In *Interspeech 2024*, pages 4593–4597, Kos, Greece.
- Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. [Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias](#). *Proceedings of the National Academy of Sciences*, 120(51).
- Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. 2022. [Probing for incremental parse states in autoregressive language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2801–2813, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhिलाषा Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. [AudioChatLlama: Towards general-purpose speech abilities for LLMs](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, Mexico City, Mexico. Association for Computational Linguistics.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. [Neural language models as psycholinguistic subjects: Representations of syntactic state](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. [A survey of adversarial defenses and robustness in nlp](#). *ACM Comput. Surv.*, 55(14s).
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. [On the robustness of self-attentive models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, Florence, Italy. Association for Computational Linguistics.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. [A systematic assessment of syntactic generalization in neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.
- Jennifer Hu and Roger Levy. 2023. [Prompting is not a substitute for probability measurements in large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.
- Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. [Language models align with human judgments on key grammatical constructions](#). *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
- Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. [Elements of world knowledge \(ewok\): A cognition-inspired framework for evaluating basic world knowledge in language models](#). *arXiv preprint*.
- Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024a. [Scope ambiguities in large language models](#). *Transactions of the Association for Computational Linguistics*, 12:738–754.
- Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024b. [Scope ambiguities in large language models](#). *Transactions of the Association for Computational Linguistics*, 12:738–754.
- Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emanuele Chersoni, Jingyuan Selena She, Zawad

- Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. [Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely](#). *Cognitive Science*, 47(11):e13386.
- Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. 2024. [Do llms understand ambiguity in text? a case study in open-world question answering](#). In *2024 IEEE International Conference on Big Data (Big-Data)*, pages 7485–7490.
- Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. [Systematic literature review: Computational approaches for humour style classification](#). *arXiv preprint*.
- Margaret Li and Julian Michael. 2022. [Overconfidence in the face of ambiguity with adversarial data](#). In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pages 30–40, Seattle, WA. Association for Computational Linguistics.
- Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022. [VISA: An ambiguous subtitles dataset for visual scene-aware machine translation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6735–6743, Marseille, France. European Language Resources Association.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. [We’re afraid language models aren’t modeling ambiguity](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.
- James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. [Can Peanuts Fall in Love with Distributional Semantics?](#) In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, Sydney, Australia. Cognitive Science Society.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Kanishka Misra, Allyson Ettinger, and Kyle Mahowald. 2024. [Experimental contexts can facilitate robust semantic property inference in language models, but inconsistently](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12342–12355, Miami, Florida, USA. Association for Computational Linguistics.
- Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. [Homophone disambiguation reveals patterns of context mixing in speech transformers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.
- Charles Morris. 1946. *Signs Language and Behavior*. Prentice-Hall, New York, USA.
- Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2022. [Probing for labeled dependency trees](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7711–7726, Dublin, Ireland. Association for Computational Linguistics.
- Mante S Nieuwland and Jos JA Van Berkum. 2006. [When Peanuts Fall in Love: N400 Evidence for the Power of Discourse](#). *Journal of Cognitive Neuroscience*, 18(7):1098–1111.
- Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. [The communicative function of ambiguity in language](#). *Cognition*, 122(3):280–291.
- Charlotte Pouw, Marianne de Heer Kloots, Afra Alishahi, and Willem Zuidema. 2024. [Perception of phonological assimilation by neural speech recognition models](#). *Computational Linguistics*, 50(3):1557–1585.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Walid S. Saba and Jean-Pierre Corriveau. 2001. [Plausible reasoning and the resolution of quantifier scope ambiguities](#). *Studia Logica: An International Journal for Symbolic Logic*, 67(2):271–289.
- Bushra Sabir, M. Ali Babar, and Sharif Abuadbbba. 2023. [Interpretability and transparency-driven detection and transformation of textual adversarial examples \(it-dt\)](#). *arXiv preprint*.
- Michael Spivey-Knowlton and Julie C Sedivy. 1995. [Resolving attachment ambiguities with multiple constraints](#). *Cognition*, 55(3):227–267.
- Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2023. [Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in VQA](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10220–10237, Toronto, Canada. Association for Computational Linguistics.

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. 2024. [Zero and few-shot semantic parsing with ambiguous inputs](#). In *The Twelfth International Conference on Learning Representations*.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. [PhonologyBench: Evaluating phonological skills of large language models](#). In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Abdul Karim Taha. 1983. [Types of syntactic ambiguity in english](#). *International Review of Applied Linguistics in Language Teaching*, 21(4):251–266.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The Benchmark of Linguistic Minimal Pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.

Vera Zaboltnina, Didier Bottineau, and Elena Boyarskaya. 2021. [Cognitive mechanisms of ambiguity resolution](#). In *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics*, pages 201–212, Cham. Springer International Publishing.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. 2024. [Text-crs: A generalized certified robustness framework against textual adversarial attacks](#). In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2920–2938.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. [Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5482–5492, Online. Association for Computational Linguistics.

Xiliang Zhu, Chia-Tien Chang, Shayna Gardiner, David Rossouw, and Jonas Robertson. 2024. [Resolving transcription ambiguity in Spanish: A hybrid acoustic-lexical system for punctuation restoration](#). In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 33–41, Malta. Association for Computational Linguistics.

A Motivation

The motivation behind creating AmbAdv, the first adversarial ambiguity dataset, lies in the fact that ambiguity remains a persistent challenge for language models. Prior work has explored ambiguity in narrow contexts, such as natural language

inference (NLI) (Liu et al., 2023), semantic parsing (Stengel-Eskin et al., 2024), and log probabilities (Kamath et al., 2024a). These studies primarily focus on syntactic and scope ambiguity, assessing how LLMs handle ambiguous inputs in zero-shot and few-shot settings. However, previous work has not directly assessed LLMs’ ability to detect ambiguity through user-facing interactions. Since users engage with LLMs via direct prompts, it is crucial to understand how models respond to ambiguous inputs in real-world usage.

To systematically evaluate ambiguity sensitivity, we focus on three key ambiguity types: syntactic, lexical, and phonological. This selection is grounded in cognitive and linguistic theories (Zaboltnina et al., 2021), which classify ambiguity into lexical, phonological, morphological, and syntactic categories. We put a great focus on syntactic ambiguity, particularly PP attachment ambiguities, as it poses challenges for parsing and can lead to garden-path effects. Moreover, identifying between the two possible structures/readings of a syntactically ambiguous sentence often requires background world knowledge (Saba and Corriveau, 2001), which is a challenging concept for an artificial intelligence model (Ivanova et al., 2024). We included lexical ambiguity, as it is the most frequent, as word meanings are highly flexible and influenced by both linguistic and extralinguistic factors. Lastly, phonological ambiguity, though rarely explored in computational settings, is crucial in spoken language, humour, and wordplay—yet, to our knowledge, no dataset has included it. Morphological ambiguity was not included, as it was infeasible to systematically construct adversarial variations.

By focusing on these ambiguity types, we aim to broaden the scope of ambiguity evaluation and examine whether LLMs can recognize and handle the diverse sources of linguistic uncertainty that shape human communication.

A.1 Comparison with Prior Datasets

While several recent datasets have explored ambiguity in language models, AmbAdv is unique in its adversarial construction and coverage of ambiguity types. Unlike prior resources such as AmbigQA, Ambient, AMP, and Scope, which primarily focus on naturally occurring ambiguities within specific tasks like open-domain question answering or natural language inference, AmbAdv systematically introduces controlled perturbations—such

Dataset	Size	Ambiguity Types	Use Case	Adversarial
AmbigQA (Min et al., 2020)	14,042	Referential	Open-domain QA	✗
Ambient (Liu et al., 2023)	1,645	Lexical, Syntactic	NLI, Disambiguation	✗
AMP (Stengel-Eskin et al., 2024)	NA-Synthetic	Lexical, Syntactic	Semantic Parsing	✗
Scope (Kamath et al., 2024a)	1,000	Scope	Human Preference Analysis	✗
AmbAdv (Ours)	1,529	Syntactic, Lexical, Phonological	Ambiguity Identification, Disambiguation	✓

Table 4: Comparison of AmbAdv with related ambiguity datasets in terms of size, ambiguity types, intended use case, and adversarial construction.

as synonym replacements, random substitutions, and rhyming alterations—to create ambiguous and unambiguous sentence pairs. This adversarial approach is designed to rigorously evaluate model robustness in handling ambiguity. Additionally, AmbAdv is the first to incorporate phonological ambiguity (i.e., homonyms) into its evaluation framework, addressing a previously underexplored area in ambiguity research. By centering on the task of ambiguity identification, AmbAdv challenges models to detect and interpret ambiguous inputs without external context, thereby complementing existing resources and providing a valuable benchmark for advancing research in ambiguity detection and resolution within NLP systems. Table 4 provides a summary that can help position our dataset relative to prior work.

B Dataset Details

The dataset was curated by the authors of the paper after collecting linguistic ambiguity sentences from various online linguistic textbooks and publicly available datasets (Taha, 1983; Liu et al., 2023; Stengel-Eskin et al., 2024). For syntactic ambiguity, we generated a total of 1,097 sentences, with sentence variants for each manipulation type and part of speech. We decided to provide multiple variations for the different parts of the sentence, as the set of sentences has a similar structure, and we wanted to reflect how ambiguity arises from a network of heterogeneous participants—agents, subjects, and objects—each with varying roles (Zabotkina et al., 2021). For lexical and phonological sentences, we created 4 examples per manipulation type, resulting in a total of 224 sentences for lexical ambiguity and 208 for phonological ambiguity (we did not have a different word order in phonological ambiguity because the homonyms changed).

Many researchers have explored word substitution as a form of adversarial attack in NLP (Zhou et al., 2021; Bespalov et al., 2023; Sabir et al.,

2023; Zhang et al., 2024). Building on this foundation, we selected synonym replacement, word order changes, and random perturbations as core transformation types, as these are commonly used in adversarial defense benchmarks and robustness evaluations (Goyal et al., 2023). To extend beyond traditional lexical and syntactic manipulations, we introduced a novel rhyme-based perturbation, motivated by the underexplored area of phonological ambiguity. This choice is further supported by the growing interest in phonological effects in LLM outputs (Fathullah et al., 2024; Suvarna et al., 2024), particularly as models increasingly integrate multimodal capabilities. By incorporating this diverse set of perturbations, our study aims to investigate a broader spectrum of ambiguity types and assess model robustness across both well-studied and emerging linguistic phenomena.

Rhyme substitutions explore phonological similarity, an underexplored adversarial strategy in NLP (Suvarna et al., 2024). These are particularly relevant given the increasing attention to phonological effects in LLM outputs, especially in spoken or dialogue contexts (Fathullah et al., 2024).

Table 5 presents the complete set of base sentences used for syntactic ambiguity. Table 6 lists all base sentences for phonological ambiguity, and Table 7 provides the full set for lexical ambiguity.

No.	Sentence
1	The man saw the woman with the telescope.
2	She fed her cat food.
3	Harry loves his pet turtle more than his wife.
4	The captain ordered the old men and women of the ship.
5	I saw a dog in my pyjamas.
6	An enraged cow injured a farmer with an ax.
7	The hospital is being sued by six foot doctors.
8	Helen got lunch ready for her daughter wearing a summer dress.

Table 5: Set of syntactically ambiguous base sentences.

To ensure consistency in our annotations, we applied distinct validation strategies for each ambiguity type. Lexical ambiguity was identified through dictionary lookup, confirming that a word had multiple meanings depending on context. Phonological

No.	Sentence Pair
1	It's not easy to wreck a nice beach. / It's not easy to recognize speech.
2	I saw a sea monster. / I saw a seam on stir.
3	She sells seashells by the seashore. / She sells sea shells buy the sea sure.
4	Whale meat again. / We'll meet again.
5	I saw a bear in the forest. / Eye sore a bare inn the for rest.
6	He took a nice cold shower after his date. / He took an ice cold shower after his date.
7	The stuffy nose is annoying. / The stuff he knows is annoying.
8	He couldn't wait to leave the mall. / He couldn't wait to leave them all.
9	The good can decay many ways. / The good candy came anyways.
10	She can't bear to lose the race. / She can't bear to lose their ace.
11	He's a man of many talents. / He's a man of mini talents.
12	I hurt myself with the four candles. / I hurt myself with the fork handles.
13	I ordered pepperoni pizza. / I ordered pepper only pizza.
14	They wanted to explore the ancient ruins. / They wanted to explore the agent's ruins.
15	I love you. / Aisle of view.
16	Caesar salad. / Seize her salad.

Table 6: Set of phonologically ambiguous sentence pairs.

No.	Sentence
1	Your boss is a funny man.
2	The speaker is at the front of the room.
3	He's not very well off.
4	John and Anna are married.
5	It is my belief that the earth is round.
6	She is looking for a match.
7	Give me a ring.
8	Show me the light feathers.
9	We saw her duck.
10	I'm going to take a break from studying.
11	Alice and Jon disagreed.
12	Give me the bat!
13	What you said is insane.
14	Yesterday I went to the bank.
15	I can't find the glasses.
16	He didn't see the picture of the disaster.

Table 7: Set of lexically ambiguous sentences.

ambiguity was validated using IPA transcriptions from authoritative linguistic resources to verify homophony or rhyming patterns. In contrast, syntactic ambiguity was treated differently: all syntactic examples were constructed to allow multiple structural interpretations. However, we observed that many such cases could be pragmatically disambiguated based on world knowledge. For instance, in the sentence “*The man saw the woman with the dress*,” although the syntax permits multiple parses, common sense rules out interpreting the dress as a visual instrument. Based on this reasoning, we labeled sentences with random or rhyme-based substitutions as *unambiguous*, since their ambiguity was not plausible under typical interpretive contexts.

All annotators were recruited via the Prolific crowdsourcing platform⁷ and compensated fairly at a rate of \$18 per hour. The only eligibility criterion was being a native English speaker, with no restrictions on geographic location. Annotators received task instructions closely aligned with

⁷<https://www.prolific.com/>

the default prompts used in our experiments (see Appendix C.1 for details), ensuring consistency between model and human evaluations.

We also present in Figure 4 a distribution of the ambiguous or non-ambiguous sentences comparing the *Gold Label* as annotated in the datasets, and the model predictions.

C Experiments

C.1 Prompt Templates

We used a total of 8 different templates for ambiguity identification: 2 default templates -one asking directly if the sentence is ambiguous and one asking indirectly if the sentence has two interpretations-, and 6 templates with a binary word alteration between true, false, yes, no, right, and wrong. We also used 2 templates for ambiguity disambiguation, which are modified versions of the default templates. We provide indicative examples of the Llama with the llama chat template in Figures 5-13. All templates and model outputs will be available upon releasing our codebase.

C.2 Results across Prompts

We present the results across the different prompt templates in Table 8. Overall we observe that models perform better on different prompts, and there does not seem to be an optimal prompt. The variations may be marginal in some cases, suggesting that most models we examined are not prompt-sensitive. The only model that seems to have a preference for a prompt is Gemma, which seems to prefer the default2 prompt asking directly if a sentence is ambiguous.

We also provide a random sample of responses from Qwen-2 and Llama-3 using the disambiguation prompt in Figure 9 in Tables 9 and 10. below. The full model responses will be available in our codebase.

C.3 Results Across Syntactic Roles

While creating the dataset for syntactic ambiguity, we carefully altered the sentences according to the manipulations described in Table 2 by changing specific sentence components. In particular, we modified the subject, object, the prepositional phrase (if any), and all components in the set of our original sentences. This division reflects the fact that ambiguity arises from a network of heterogeneous participants—agents, subjects, and objects—each with varying roles, reflexive awareness,

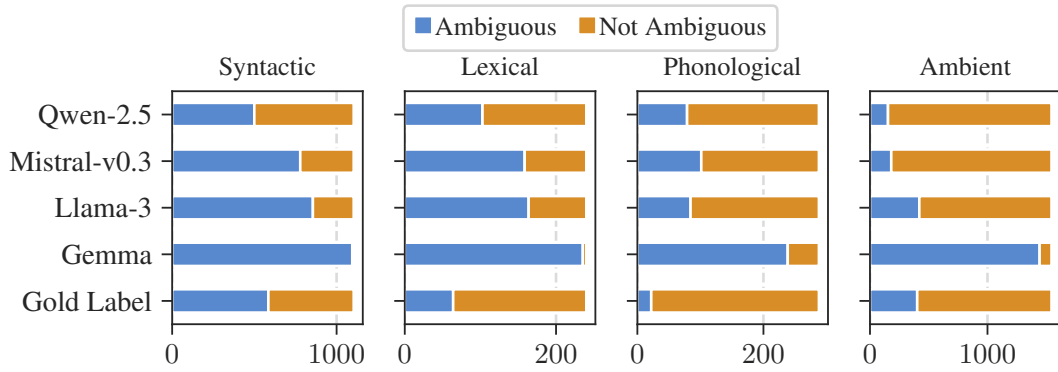


Figure 4: Comparison of the ambiguous and non-ambiguous count of sentences predicted by language models against the 'gold standard label'.

Model	Default1	Default2	False	No	Right	True	Wrong	Yes	AVG
Qwen-2.5	0.409	0.409	0.473	0.479	0.433	0.444	0.426	0.414	0.436
Mistral-v0.3	0.503	0.378	0.356	0.306	0.322	0.304	0.368	0.340	0.360
Llama-3	0.511	0.486	0.440	0.473	0.447	0.450	0.439	0.447	0.462
Gemma	0.530	0.489	0.531	0.531	0.531	0.531	0.531	0.531	0.525

(a) Syntactic Ambiguity

Model	Default1	Default2	False	No	Right	True	Wrong	Yes	AVG
Qwen-2.5	0.354	0.354	0.608	0.588	0.550	0.617	0.542	0.312	0.491
Mistral-v0.3	0.263	0.287	0.388	0.508	0.458	0.458	0.438	0.454	0.407
Llama-3	0.271	0.438	0.400	0.375	0.375	0.404	0.333	0.287	0.360
Gemma	0.267	0.283	0.267	0.267	0.267	0.267	0.267	0.263	0.268

(b) Lexical Ambiguity

Model	Default1	Default2	False	No	Right	True	Wrong	Yes	AVG
Qwen-2.5	0.795	0.753	0.753	0.747	0.802	0.840	0.764	0.792	0.781
Mistral-v0.3	0.646	0.177	0.681	0.733	0.726	0.819	0.715	0.681	0.647
Llama-3	0.663	0.788	0.809	0.792	0.670	0.670	0.701	0.486	0.697
Gemma	0.076	0.198	0.198	0.319	0.306	0.212	0.323	0.347	0.247

(c) Phonetic Ambiguity

Model	Default1	Default2	False	No	Right	True	Wrong	Yes	AVG
Qwen-2.5	0.654	0.654	0.706	0.700	0.712	0.711	0.707	0.700	0.693
Mistral-v0.3	0.570	0.702	0.712	0.735	0.724	0.733	0.718	0.727	0.703
Llama-3	0.524	0.729	0.672	0.631	0.673	0.672	0.632	0.551	0.636
Gemma	0.261	0.494	0.265	0.265	0.271	0.271	0.267	0.299	0.299

(d) Ambient Premises

Table 8: Accuracy Results for Different Ambiguity Sets


```

<|start_header_id|>system<|end_header_id|>

Considering real-world knowledge, could this sentence be interpreted in more than
one way?

{{ sentence }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>

```

Figure 5: Default1 prompt used for identifying syntactic and lexical ambiguities.

```

<|start_header_id|>system<|end_header_id|>

Based on general world knowledge, is the following sentence ambiguous?

{{ sentence }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>

```

Figure 6: Default2 prompt used for identifying syntactic and lexical ambiguities.

```

<|start_header_id|>system<|end_header_id|>

Based on general world knowledge, is the following sentence ambiguous?
Reply with True or False.

{{ sentence }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>

```

Figure 7: Example of the True binary type alteration prompt for identifying syntactic and lexical ambiguities.

and intentionality (Zabotkina et al., 2021).

After prompting our models, we then examined whether there were any salient differences in error rate across syntactic roles. We report the results of this experiment in Figure 14. Overall, we find that the results are relatively stable across syntactic roles and no significant patterns could be identified.

D Representational Analysis

We also provide further insights from the representational analysis across more layers. We present examples from the initial (layer 0), early (layer 6), middle (layer 12), and late layers (layer 24) across all token roles. We show PCA projections for Llama-3 in Figure 15 and Gemma in Figure 16 across layers and token types. We extract representations of the final end of prompt sentence represen-

```

<|start_header_id|>system<|end_header_id|>

Based on general world knowledge, is the following sentence ambiguous?
Reply with False or True.

{{ sentence }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>

```

Figure 8: Example of the False binary type alteration prompt for identifying syntactic and lexical ambiguities.

```

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Based on general world knowledge, is the following sentence ambiguous? If yes,
provide the two disambiguations of the sentence separated by a comma.

{{ sentence }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 9: Example of the prompt used for the disambiguation task.

```

<|start_header_id|>system<|end_header_id|>

Could these two sentences sound alike but have different meanings?

Sentence 1: {{ sentence1 }}
Sentence 2: {{ sentence2 }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>

```

Figure 10: Default1 prompt used for identifying phonological ambiguity.

tation (eos), the first verb, first noun, and last noun. Across models and token types, representations extracted at the first layer, which are not yet contextualized via the encoder module, do not allow for decoding ambiguity, and no meaningful clustering structure has emerged yet. At early layer 6, the last noun already achieves remarkably high probe accuracies in both Llama-3 (0.83) and Gemma (0.93), which consistently increase towards later layers.

Similarly, but at an overall lower level, the first verb and first noun, also achieve higher probe accuracies of 0.6 – 0.69, but plateau from middle layers on for both models. Interestingly, the eos probe accuracies remain low in early layers at 0.45 (Llama-3) and 0.55 (Gemma), but increase from then on and achieve high scores in late layers of around 0.9. We furthermore observe a clustering of samples around the original sentence examples

```
<|start_header_id|>system<|end_header_id|>

Are these two sentences phonologically ambiguous?

Sentence 1: {{ sentence1 }}
Sentence 2: {{ sentence2 }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>
```

Figure 11: Default2 prompt used for identifying phonological ambiguity.

```
<|start_header_id|>system<|end_header_id|>

Are these two sentences phonologically ambiguous? Reply with True or False.

Sentence 1: {{ sentence1 }}
Sentence 2: {{ sentence2 }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>
```

Figure 12: Example of the True binary type alteration prompt for identifying phonological ambiguity.

```
<|start_header_id|>system<|end_header_id|>

Are these two sentences phonologically ambiguous? Reply with False or True.

Sentence 1: {{ sentence1 }}
Sentence 2: {{ sentence2 }}

Format your response according to the following JSON schema:
{{ schema }}
<|eot_id|>
```

Figure 13: Example of the False binary type alteration prompt for identifying phonological ambiguity.

used to build adversaries and variations contained in our AmbAdv dataset, e.g., for the first verb, this appears across all layers in both models, hinting at the model’s limited ability to accurately disambiguate sentence meaning from the verb alone. Our results further hint at a special role of the last noun, as the model appears to distinguish between ambiguous and non-ambiguous instances through the last noun representation. Importantly, several token

types outperform the behavioural model evaluation, while our representational probe analysis of the eos token highlights that relevant information for accurate inferences is available, but can not be reliably used by the model for generating correct answers. In conclusion, our analysis highlights the evolving role of token representations across layers, with a particular emphasis on the last noun as a central factor in disambiguating sentence meaning.

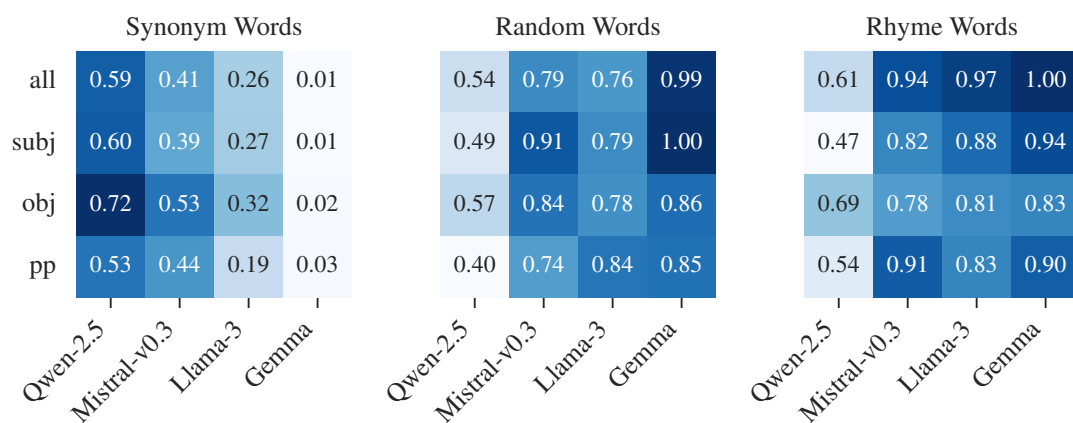


Figure 14: Results across models after zero-shot prompting, showing the error rate of the different synonym, random, and rhyme word modifications. Values closer to 1 indicate a higher error rate. We denote the substitution of the subject components of a sentence as *subj*, the object components as *obj*, and the prepositional phrases as *pp*. Lastly, when we substitute more than two words we refer to this manipulation as *all*.

Text	Is Ambiguous	Disambiguations
The man saw the woman with the telescope.	yes	['The man used a telescope to see the woman, The man saw the woman who was holding a telescope']
The woman saw the man with the telescope.	yes	['The woman used a telescope to see the man, The woman saw the man who was holding a telescope']
The guy saw the girl with the telescope.	yes	['The man used a telescope to see the woman, The man saw the woman who was using a telescope']
The man saw the woman with the microscope.	yes	['The man used the microscope to see the woman, The man saw the woman who was using a microscope']
The man saw the woman with the monocular.	yes	['The man used a monocular to see the woman, The man saw the woman who was holding a monocular']
The man saw the woman with the camera.	yes	['The man saw the woman who was holding a camera, The man saw the woman using a camera to take a picture.']
The man saw the woman with the book.	yes	['The man saw the woman holding a book, The man saw the woman who was a writer of a book']
The man saw the woman with the dress.	yes	['The man saw the woman wearing the dress, The man saw the woman who was with the dress (possibly holding it or in possession of it).']
The man saw the woman with the suitcase.	yes	['The man saw a woman who was carrying a suitcase, The man saw a woman with a suitcase (near her)']
The mug saw the woman with the telescope.	yes	['A mug witnessed a woman using a telescope, A person named Mug observed a woman with a telescope.']
The telescope saw the man with the woman.	yes	['The telescope was used to observe a man and a woman together, The telescope had a man and a woman looking through it at the same time.']
The detective saw the woman with the telescope.	yes	['The detective saw the woman using a telescope, The detective saw the woman near a telescope']
The teacher saw the woman with the telescope.	yes	['The teacher used a telescope to see the woman, The teacher saw the woman who was carrying a telescope']
The mother saw the woman with the telescope.	yes	['The mother used a telescope to see the woman, The mother saw the woman holding a telescope']
The father saw the woman with the telescope.	yes	['The father saw a woman who was using a telescope, The father saw a woman with a telescope, meaning a woman who is associated with or has a connection to a telescope']
The artist saw the woman with the telescope.	yes	['The artist saw a woman who was using a telescope, The artist saw a woman who was looking at something through a telescope']
The explorer saw the woman with the telescope.	yes	['The explorer saw a woman who was using a telescope, The explorer saw a woman with a telescope, i.e. the woman was in possession of the telescope']
The man saw the bridge with the telescope.	yes	['The man saw the bridge through the telescope, The man saw the bridge that had a telescope on it']
The man saw the eagle with the telescope.	yes	['The man saw the eagle through the telescope, The man saw the eagle that had a telescope on it']
The man saw the weather with the telescope.	yes	['The man used a telescope to observe the weather, The man saw the weather conditions through the telescope']
The man saw the beer with the telescope.	yes	['The man saw the beer using the telescope, The man looked at the beer with the aid of a telescope']
The man saw the bee with the telescope.	yes	['The man saw the bee with his own eyes, The man used the telescope to see the bee']
Jenny saw Alex with the camera.	yes	[Jenny saw Alex holding a camera, Jenny saw Alex use a camera to take a picture']
The boy saw the girl with the glasses.	yes	['The boy saw the girl wearing glasses, The girl was accompanied by a boy who was wearing glasses']
The biologist saw the student with the telescope.	yes	['The biologist saw the student using the telescope, The biologist saw the student with (near or holding) a telescope']

Table 9: Ambiguity predictions and their disambiguations according to Qwen-2.5

Text	Is Ambiguous	Disambiguations
The man saw the woman with the telescope.	yes	['The man saw the woman while using the telescope, The man saw the woman who was looking through the telescope.']
The woman saw the man with the telescope.	yes	['The woman saw the man who was using the telescope, The woman saw the man with the telescope (i.e., the man was in her presence, possibly holding the telescope)']
The guy saw the girl with the telescope.	yes	['The man saw the woman who was using the telescope, The guy saw the woman with a telescope, observing or accompanying her as she used it.']
The man saw the woman with the microscope.	yes	['The man saw the woman who was using the microscope, The man saw the woman holding the microscope.']
The man saw the woman with the monocular.	yes	['The man saw a woman who was using a monocular, The man saw a woman with a physical characteristic of having a monocular (e.g. one eye) or with a monocular as a possession.']
The man saw the woman with the camera.	yes	['The man saw the woman who was holding the camera, The man saw the woman while the man was using the camera.']
The man saw the woman with the book.	yes	['The man saw the woman while she was holding the book, The man saw the woman and noticed that she had the book.']
The man saw the woman with the dress.	yes	['The man saw the woman who was wearing the dress, The man saw the woman with the person who was wearing the dress.']
The man saw the woman with the suitcase.	yes	['The man saw the woman who was carrying the suitcase, The man saw the woman and then saw the suitcase.']
The telescope saw the man with the woman.	yes	['The telescope saw the man who was with the woman, The telescope saw the man while the woman was with it.']
The mug saw the woman with the telescope.	yes	['The mug, a household item, is being used as a metaphor to describe something that saw the woman with the telescope, possibly implying that the mug has a unique perspective or insight. The mug, a household item, is literally being used to describe the object that saw the woman with the telescope, possibly implying that the mug has eyes or is being used as a tool to observe the woman with the telescope.']
The detective saw the woman with the telescope.	yes	['The detective saw a woman who was using a telescope, The detective saw a woman who was being looked at with a telescope.']
The teacher saw the woman with the telescope.	yes	['The teacher saw a woman who was using a telescope, The teacher saw a woman who was being looked at through a telescope.']
The mother saw the woman with the telescope.	yes	['The mother saw the woman who was using the telescope, The mother saw the woman with the telescope, but the woman was not using it.']
The father saw the woman with the telescope.	yes	['The father saw a woman who was using a telescope; The father saw a woman with a telescope, meaning a woman who is associated with or has a connection to a telescope.']
The artist saw the woman with the telescope.	yes	['The artist saw a woman who was using a telescope, The artist saw a woman who was looking at something through a telescope']
The explorer saw the woman with the telescope.	yes	['The explorer saw a woman who was using a telescope, The explorer saw a woman with a telescope, i.e. the woman was in possession of the telescope.']
The man saw the weather with the telescope.	yes	['The man used a telescope to observe the weather, The man saw the weather conditions through the telescope.']
The man saw the beer with the telescope.	yes	['The man saw the beer using the telescope, The man looked at the beer with the aid of a telescope.']
The man saw the bee with the telescope.	yes	['The man saw the bee with his own eyes, The man used the telescope to see the bee.']
Jenny saw Alex with the camera.	yes	['Jenny saw Alex with the camera, meaning Jenny used the camera to take a picture of Alex; Jenny saw Alex who was holding the camera.']
The boy saw the girl with the glasses.	yes	['The boy saw a girl who was wearing glasses, The boy saw a girl who was with someone wearing glasses.']
The biologist saw the student with the telescope.	yes	['The biologist was using a telescope to look at the student, The biologist saw a student who was using a telescope.']

Table 10: Ambiguity predictions and their disambiguations according to Llama 3

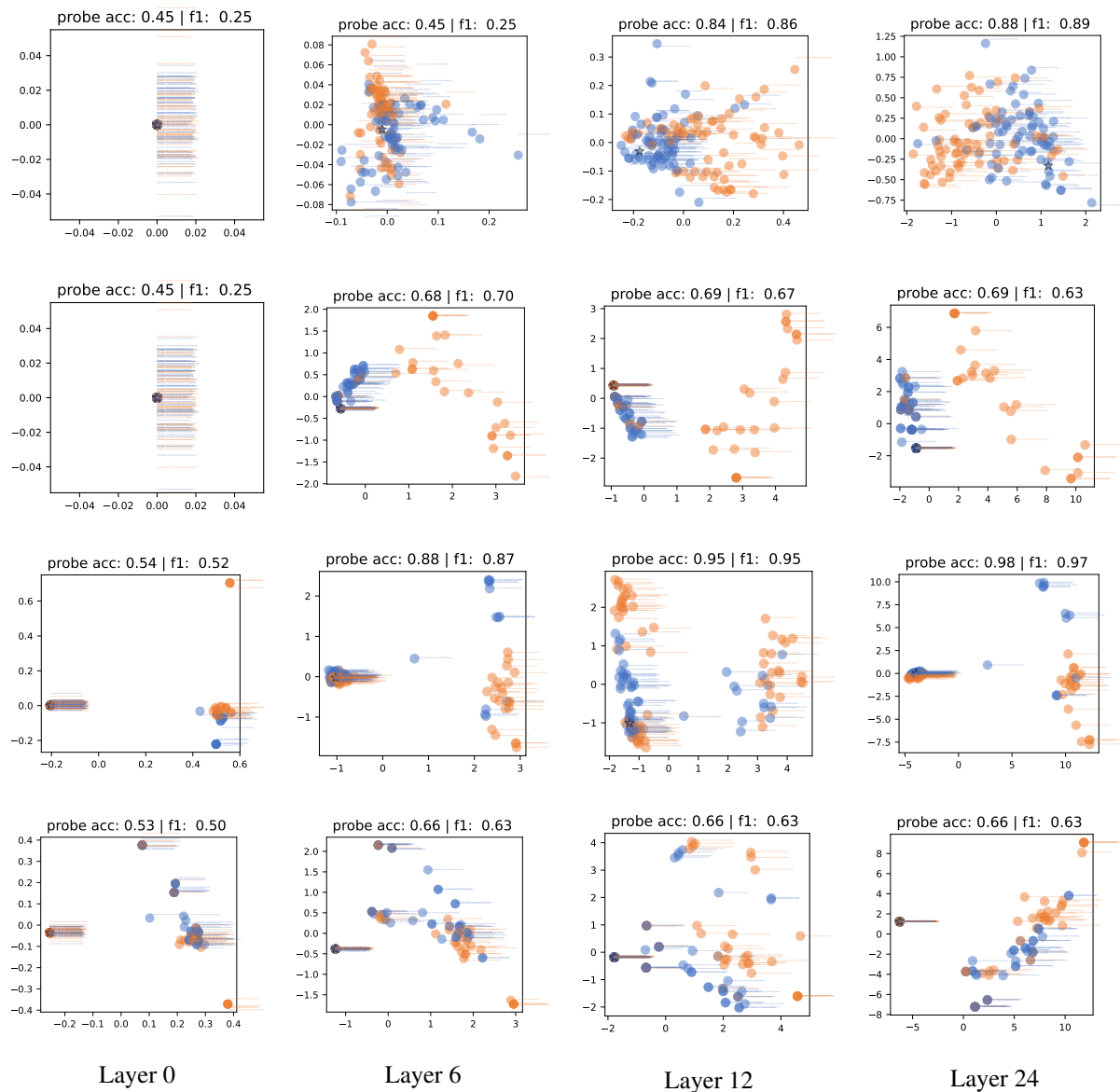


Figure 15: PCA projections extracted for different token roles (eos, verb, last noun, first noun) across layers for Llama 3. Blue/orange indicate ambiguous/not ambiguous sentences.

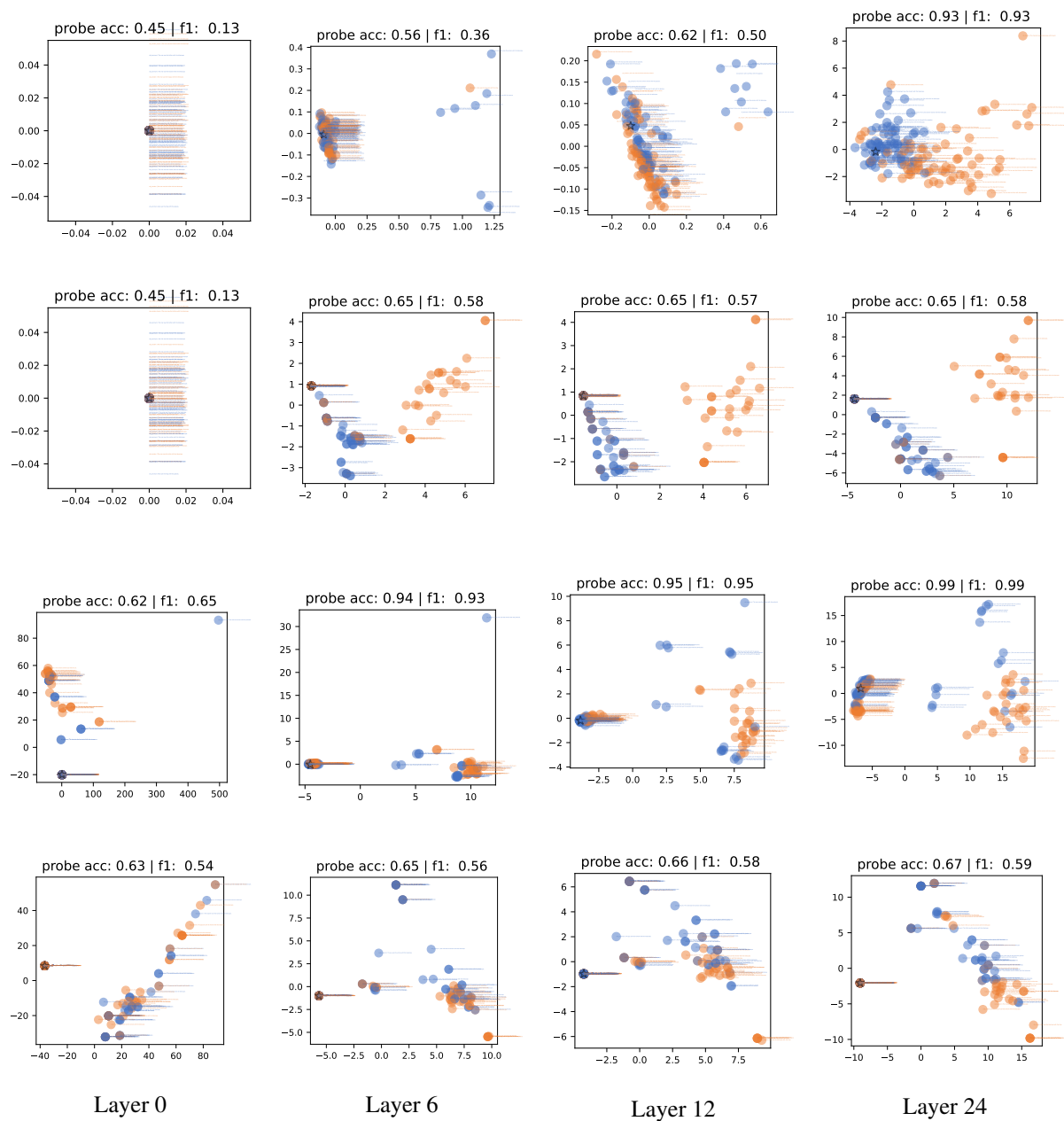


Figure 16: PCA projections extracted for different token roles (eos, verb, last noun, first noun) across layers for Gemma. Blue/orange indicate ambiguous/not ambiguous sentences.