Convergence
A Journal for Young Researchers

ORIGINAL RESEARCH

Received: 8-28-25; Accepted: 10-31-25;
Managing Editor: Jonas Katona;
Action Editor: Jonas Katona

Ambiguous by Design: Testing LLMs on Polysemy, Homonymy, and Underspecification

Myra Arora

Bal Bharati Public School, Rohini, Delhi, India

Abstract

Large Language Models (LLMs) continue to show remarkable fluency and versatility across a range of tasks. However, their performance is heavily influenced by the structure and clarity of the prompts they receive. This study investigates a critical but underexplored axis of prompt design: lexical ambiguity. We construct a controlled dataset of prompt pairs, each containing an ambiguous query and its corresponding disambiguated version, and evaluate outputs from multiple state-of-the-art LLMs (GPT-4, Claude 4 Sonnet, and Gemini 2.5 Pro). Using a combination of semantic similarity (BERTScore), hallucination detection (FactCC), and confidence calibration metrics (hedging frequency, assertiveness index), we quantify the effects of ambiguity on output quality. Our results show that ambiguous prompts significantly increase the likelihood of misinterpretation and factual hallucination, and often reduce the model's confidence calibration. Moreover, we observe model-specific variance in ambiguity robustness, suggesting that prompt sensitivity is architecture-dependent. These findings illuminate a critical failure point in prompt–response alignment and offer empirical insights for improving prompt engineering practices. This work presents one of the first empirical investigations of lexical ambiguity in LLM interaction, with implications for prompt safety, automated reasoning, and model interpretability.

Keywords: polysemy, lexical ambiguity, prompt engineering, large language models (LLMs), artificial intelligence (AI)

1. Introduction

1.1. Context and Importance

Large Language Models (LLMs) are deep neural models that have been trained on huge text datasets to predict the next token in a sequence. Through this autoregressive objective, they implicitly learn syntactic, semantic, and even pragmatic patterns of natural language. Given textual instructions or questions as input, they make predictions conditioned on the input. The formulation of such prompts has a determinative impact on response quality, style, and factuality, so prompt engineering becomes a primary aspect of both testing and implementation.

Contemporary LLMs contain several guardrails, including content filtering, reinforcement learning from human feedback (RLHF), and safety fine-tuning, to down-regulate unsafe or hallucinatory responses. These protective measures, though, do not completely eradicate hallucinations, especially under ambiguous or polysemous situations. At inference time, a number of parameters that can be controlled condition generation behavior. Foremost among these is temperature, which controls the level of randomness in token selection: lower temperatures (e.g., 0-0.3) produce more deterministic and repetitive output, while higher settings (e.g., 0.8-1.0) foster diversity and imagination but sacrifice factual coherence. In this study, all models were evaluated at a fixed temperature of 0.7, a widely used mid-range value that balances coherence with variability and enables fair cross-model comparison.

LLMs such as GPT-4, Claude, and Gemini have become foundational tools in natural language processing, powering applications in education, customer support, scientific writing, and legal aid. Their apparent fluency, sophistication, and contextual sensitivity enable them to scale across domains. But this smoothness hides an essential brittleness: LLMs are disproportionately prompt-dependent. As demonstrated in previous research (Gonen et al., 2023; Melamed et al., 2025), small variations in prompt syntax (e.g., syntax or length) can create disproportionate changes in output correctness and reasoning. In high-stakes scenarios, this brittleness is a severe shortcoming.

Small differences in word order, lexical preference, or tone can produce out-of-proportion changes in output content, reasoning path, or factuality. This is because LLMs lack a grounded understanding of human intent; instead, they work by maximizing the probability of the next token according to learned statistical patterns. This two-sidedness—greatness and fragility—keeps prompting as both their strongest asset and their most troublesome point of failure.

1.2. Polysemy, Homonymy, and Semantic Underspecification

In linguistic semantics, polysemy refers to a single word having multiple related meanings (e.g., “paper” as a material vs. “paper” as an academic article). Homonymy, in contrast, involves words that share the same form but have entirely unrelated meanings (e.g., “bank” as a financial institution vs. “bank” of a river). Semantic underspecification occurs when a word or phrase lacks sufficient context to determine a single, precise interpretation (e.g., “Alex saw her duck”, where “duck” could mean an action or an animal).

These distinctions are central to evaluating how LLMs process lexical ambiguity, since each type of ambiguity challenges different levels of representational reasoning.

1.3. The Problem: Lexical Ambiguity in Prompts

One of the most pernicious forms of prompt brittleness arises from lexical ambiguity: when a word or phrase has multiple plausible meanings. Words like pitch, bass, or bark can be used to describe completely different things in different contexts. Even in business usage scenarios, prompts such as "describe the war," "speak about light consumption," or "explain the role of a bank" are routine—and semantically underspecified.

Humans routinely disambiguate such expressions through shared cultural knowledge, pragmatic cues, or situational context. LLMs do not reason like this, though. They instead depend on the statistical token distribution in their training set.

When presented with an ambiguous prompt, they do not necessarily recover the intended meaning; rather, they stake out one meaning with strong confidence, usually inserting hallucinations along the way (Liu et al., 2023; Su et al., 2024).

Researchers have also begun developing benchmarks explicitly targeting ambiguity. The AmbiEnt benchmark (introduced by Liu et al., 2023) comprises human-curated ambiguous sentences designed to test LLM interpretive accuracy. Their results showed that GPT-4 selected the wrong interpretation in roughly 68% of cases. Similarly, the AmbiBench framework (proposed by Tamkin et al., 2022) extends this effort by evaluating task-level ambiguity across multiple datasets, revealing that both human and model judgments diverge systematically. These benchmarks are crucial steps toward operationalizing ambiguity, though they remain limited to isolated sentence-level contexts rather than natural prompts.

1.4. Current Gaps in Literature

While ambiguity is highly significant in human communication, very few empirical studies investigate how LLMs react to ambiguous input in open-ended generation tasks. The majority of prompt engineering studies examine formatting (Gonen et al., 2023), length (Melamed et al., 2025), or instruction fine-tuning (Kim et al., 2024). The effect of the meaning ambiguity of prompts, particularly lexical and polysemous ambiguity, has largely been neglected.

Additionally, current research on hallucination (Huang et al., 2023; Li et al., 2024) takes for granted that the input is well-specified and that hallucination is a model failure property. However, if the prompt itself is underspecified, how can we isolate hallucination from misinterpretation? This is an open issue that makes it hard to determine whether LLMs are "making things up" or merely misunderstanding under-specified instructions. To date, no major study has directly examined how lexical ambiguity in user prompts affects hallucination rate, interpretive consistency, and factual alignment across different LLMs.

1.5. Research Questions and Hypotheses

To fill this gap, this paper investigates the impact of lexical ambiguity in prompts on the behavior of LLMs. The central research questions are:

- RQ1: How does lexical ambiguity in prompts affect the factual accuracy of LLM responses?
- RQ2: Do ambiguous prompts increase hallucination frequency or cause divergence in interpretation across models?
- RQ3: Are larger, instruction-tuned models (e.g., GPT-4, Claude) more robust to ambiguity than smaller models (e.g., GPT-3.5)?

We hypothesize that:

- Ambiguous prompts will yield significantly lower semantic similarity scores between intended and generated outputs compared to clarified versions.
- Hallucination rates will increase in the presence of lexical ambiguity.
- Larger and more recently fine-tuned models will hedge or defer more often, demonstrating partial ambiguity awareness.

1.6. Contribution and Significance

This paper presents a controlled benchmark for ambiguous and disambiguated prompt pairs over three lexical ambiguity types: polysemy, homonymy, and underspecification. We measure the effect of prompt ambiguity on hallucination rate, semantic drift, and interpretive divergence with automatic metrics such as BERTScore, hallucination detectors, and hedging analysis based on outputs from GPT-4, Gemini 2.5 Pro, and Claude 4 Sonnet.

Our work makes three core contributions:

1. It provides one of the first empirical analyses of prompt-level ambiguity in natural LLM use cases.
2. It introduces a diagnostic methodology for testing LLM robustness to lexical ambiguity.
3. It offers practical implications for safe prompt design and ambiguity-aware model alignment, especially in domains like education, law, and healthcare, where misinterpretation may have high consequences.

2. Literature Review

Recent research on large language models (LLMs) has shed light on how much model behavior is contingent on the wording of prompts. Where many studies have shown LLM sensitivity to variation in syntax (Gonen et al., 2023; Melamed et al., 2025), few regard prompts as airtight bearers of intent. Such studies will vary word order, wordiness, or whitespace to probe stylistic sensitivity but rarely concern themselves with ambiguity of meaning. When a question like "Describe the pitch" is posed, the model has to first determine whether pitch is sound, sport, or persuasion: a semantic step rather than a syntactic one. This move repositions prompt engineering not just as optimization but as an exercise in semantic risk management.

In parallel with these results, researchers have started creating benchmarks that specifically aim for ambiguity. The AmbiEnt benchmark (Liu et al., 2023) proposed human-curated ambiguous sentences and found GPT-4 to misinterpret meaning in about 68% of instances. AmbiBench (Tamkin et al., 2022) broadened the study by analyzing ambiguity across different datasets and demonstrated that both humans and LLMs are uncertain, but in opposite directions. These papers together indicate that disambiguation in LLMs is not reflective of human intuition. But these benchmarks have a structural flaw: they compartmentalize ambiguity into isolated test sentences, not in natural user prompts. Ambiguity in actual use is unmarked, casual, and sloppy. Users hardly ever define their intent with any precision; they assume systems will prompt if something's unclear. LLMs, in contrast to humans, do not hesitate. They take an interpretation, typically the most common one, and advance. This disambiguation gap between benchmark ambiguity and interactional ambiguity is left uncharted.

Hallucination studies contribute another aspect to the problem of ambiguity. Current hallucination detection assumes the model understands the question and simply fabricates facts (Huang et al., 2023; Ji et al., 2023). But with ambiguous prompts, hallucination may be rooted in misinterpretation. If a model misreads “bass” as a fish instead of an instrument, the fabricated details are not arbitrary—they are contextually valid but intended-meaning invalid. This distinction is not just academic; it questions whether our hallucination metrics are wrongly penalizing models for errors in interpretation, not truthfulness (Li et al., 2024).

Closely related research investigates pragmatic constraints. Whereas humans disambiguate through shared pragmatics, LLMs never stop or seek clarification unless encouraged to do so (Kim et al., 2024). The APA framework provides a template pipeline for indicating ambiguity and seeking clarification, but it depends on multi-turn interaction and further training. Practically, models such as GPT-3.5 and Claude tend to boldly guess one meaning even when there is more than one viable interpretation. This pragmatic inflexibility means ambiguity is a trapdoor, not an opportunity for dialogue. Without clarification strategies or grounded intent modeling, LLMs mimic understanding but stumble on the most human aspect of language: doubt.

Prompting strategies that try to limit hallucination have also been explored by researchers. This work employs zero-shot prompting, in which models are given only the task description, to assess their intrinsic interpretive stability. Outside of zero-shot environments, chain-of-thought prompting (Wei et al., 2022), tree-of-thought prompting (Yao et al., 2023), and retrieval-augmented generation (RAG; Lewis et al., 2020) add reasoning steps or external grounding to reduce factual drift. Although prompt design is beyond the scope of this paper, these methods situate the wider endeavor to calibrate LLM reasoning with factual verisimilitude under conditions of vagueness.

One of the underutilized proxies for ambiguity awareness is hedging. Models that hedge, employing qualifiers such as "could," "might," or "possibly", are perhaps communicating uncertainty. Research such as Chen et al. (2023) indicates that higher hedging corresponds loosely with lower factual hallucination. However, older models tend to overcommit, and newer ones, such as GPT-4, are more reserved. Such inconsistency implies hedging could be a useful but presently unused metric for assessing a model's self-awareness in response to ambiguous input.

Together, the literature paints a picture of an arena contending with the confluence of ambiguity, hallucination, and prompting. Current research shows that ambiguity is both an error trigger and a diagnostic tool for seeing how LLMs know meaning. Still, the divergence between benchmarked ambiguity and real-world linguistic uncertainty remains: a divergence this research aims to fill.

3. Methodology

This paper assesses the impact of lexical ambiguity in prompts on the factual reliability, interpretive coherence, and epistemic confidence of responses from state-of-the-art large language models (LLMs). To systematically compare the impact of ambiguity, we developed a controlled prompt set and used uniform evaluation metrics on three top LLMs.

3.1. Prompt Dataset

The experiment relies on a manually curated prompt pair dataset of 40 pairs, each having:

- One of the ambiguous prompts, either polysemy, homonymy, or semantic underspecification (e.g., "Describe the pitch"),
- and one of its disambiguated counterparts, where the intended sense is made clear (e.g., "Describe the musical pitch").

Types of ambiguity were assigned as follows:

- Polysemous words (e.g., light, bank, bass): 16 pairs
- Homonyms (e.g., bark, bat, match): 12 pairs

- Underspecified mentions (e.g., "Explain the war," "Discuss the issue"): 12 pairs

Each question was written to accommodate more than one reasonable semantic meaning without context beyond the prompt. The simplified versions removed ambiguity explicitly with minimal added words to maintain prompt length equivalence.

3.2. Language Models

Each question pair was run on three commonly used instruction-tuned LLMs:

- GPT-4.1 (through OpenAI API)
- Claude 3 Opus (through Anthropic API)
- Gemini 2.5 Pro (through Google's VertexAI platform)

All models were given the same system prompts (if any) and tested in zero-shot mode at temperature 0.7. This value was chosen as it provides a balanced trade-off between determinism and creative variation. Lower temperatures (e.g., 0–0.3) produce rigid, highly deterministic outputs that may mask a model's sensitivity to ambiguous language. Higher temperatures (e.g., 0.9–1.0) introduce excessive randomness, which can confound performance comparisons. A temperature of 0.7 is widely used as a standard default in academic evaluations of generative models, as it preserves lexical diversity without destabilizing core semantic interpretations. Every prompt was given separately, with no follow-up interaction, to mimic a natural user experience.

3.3. LLMs Lack Pragmatic Awareness

Three fundamental measurements were employed to measure output behavior:

a) Accuracy and Semantic Fidelity

To assess if the model understood the ambiguous prompt differently, we calculated BERTScore between every ambiguous output and its corresponding clarified version. BERTScore measures semantic similarity in terms of contextual embeddings, providing a more meaningful measure of alignment than surface overlap. A lower BERTScore between unclear and clarified responses reflects semantic divergence, indicating the model understood the ambiguous prompt differently from its intended meaning.

b) Hallucination Rate

Hallucination was defined as the occurrence of unsupported or invented factual content in the output. We employed FactCC, a supervised fact-checking classifier, to identify hallucinations by comparing each model output against known ground truth or disambiguated reference outputs.

Outputs were flagged as hallucinated or non-hallucinated automatically based on classifier confidence thresholds. Hallucination rate was calculated as the percentage of ambiguous prompt outputs flagged for each of the 40 pairs per model.

c) Epistemic Confidence and Hedging

To examine the confidence calibration of the model in ambiguity, we carried out hedging analysis on:

- The number of hedging words (e.g., "might," "could," "possibly") with regex pattern matching;
- The Certainty Lexicon score: ratio of high-certainty words (e.g., "definitely," "always") to output length.

More hedging in ambiguity indicates self-knowledge of uncertainty, and constant or heightened certainty shows false confidence.

3.4. Statistical Analysis

All measurements were compared against paired t-tests between ambiguous and clarified answers by model. This enabled within-prompt comparison to filter out the effect of ambiguity on output behavior. Where relevant, model-wise ANOVA was also performed to compare differences in sensitivity across LLMs.

4. Results

On all three metrics of evaluation—semantic fidelity (BERTScore), hallucination rate (FactCC), and hedging behavior—lexical ambiguity in the prompts heavily affected model performance, albeit with differing magnitudes of degradation across models.

4.1. Semantic Fidelity (BERTScore)

Comparing ambiguous prompt outputs to their unambiguous counterparts, Gemini 2.5 Pro retained the most semantic similarity (mean BERTScore = 0.873, SD = 0.041), with coming second (0.846, SD = 0.048) and Claude 3 Opus third (0.819, SD = 0.052). Paired t-tests established a sharp decline in similarity among ambiguous and clarified answers for all three models ($p < 0.01$), with Gemini exhibiting the lowest mean reduction (−0.037), GPT-4 a medium decrease (−0.054), and Claude the highest (−0.071).

4.2. Hallucination Rate (FactCC)

FactCC analysis revealed that ambiguity increased hallucination rates for all models. Gemini demonstrated the lowest hallucination rate under ambiguity (14.5%, compared to 9.3% for clarified prompts), followed by GPT-4 (18.1% vs 11.4%) and Claude (22.7% vs 13.9%). The relative increase was smallest for Gemini (+5.2 pp) and largest for Claude (+8.8 pp), indicating greater robustness in Gemini's factual grounding.

4.3. Hedging and Confidence Calibration

Hedging analysis revealed that Gemini raised its application of hedging markers by 18.7% in ambiguity, GPT-4 by 14.2%, and Claude by 9.5%. On the other hand, certainty lexicon use fell by 11.6% for Gemini, 8.9% for GPT-4, and merely 4.1% for Claude. This indicates that Gemini not just recognized ambiguity more consistently but also modified its epistemic position better suited, while Claude overshot.

4.4. Qualitative Illustration of Hallucinations

Beyond aggregate metrics, qualitative inspection revealed systematic hallucinations arising from ambiguous prompts. To illustrate these behaviors, Table 1 presents representative examples across three ambiguity types.

Table 1: Two examples each across three ambiguity types: polysemy, homonymy, and underspecification.

Ambiguity Type	Ambiguous Question	Hallucinated Answer (excerpt)	Explanation of Hallucination
Polysemy	What is the current range of the F-15 fighter jet?	The F-15 jet is primarily used for target practice at military shooting ranges across the US and allied countries. Its primary range is located in Nevada.	Misinterprets 'range' (operational distance) as 'shooting range' (a location).
Polysemy	Please describe the head of the mentioned university.	The university's head is a complex mechanical device that contains the main printing unit for all official transcripts and documents. It was recently replaced in 2023.	Misinterprets 'head' (leader/chancellor) as a 'printer head' or a mechanical device.
Homonymy	What is the best way to catch a bass?	The clearest sound quality requires a large, specialized speaker cone made from woven carbon fiber and a low-pass filter set precisely at 180 Hz to prevent distortion.	Misinterprets "bass" (fish) as "bass" (low-frequency sound) and gives false, irrelevant audio engineering instructions.
Homonymy	Can you detail the history of the King's reign?	The horse's left leather rein was originally manufactured in Florence, Italy, in 1545, and later embroidered with gold thread during the Napoleonic Wars to indicate rank.	Misinterprets "reign" (rule) as "rein" (horse strap) and gives a false, detailed history of a physical object.
Underspecification	What were the findings of the most recent Vitamin D study?	The 2019 report by the *Foundation for Global Health* on Vitamin D found that daily supplementation increases adult cognitive function by 12% across all age groups.	Assigns "the study" to a specific, unmentioned organization and date, and provides a false, specific finding.
Underspecification	What kind of art did Lani create in the Unwanteds?	Lani, after mastering the art of the Silent Spell, secretly traveled to Warbler territory and used a single-strand gold thread to infiltrate the Quill borders, replacing a critical message scroll with a blank parchment to stall the next attack.	The model correctly identifies the character (Lani) and the fictional place (Artime) but invents a specific, plausible, yet entirely false plot event (the Silent Spell, the gold thread infiltration, the blank parchment mission) to answer the vague prompt "What did she create?"

4.5. Effect of Ambiguity Type on Model Performance

We stratified evaluation across ambiguity type (polysemy, homonymy, underspecification) and calculated per-sample BERTScore distributions for each model on an augmented dataset. Table 2 aggregates these distributions. In line with Section 5.1, averaged over the types of ambiguity, Gemini 2.5 Pro achieved the highest semantic similarity on ambiguous inputs (mean BERTScore = 0.873, SD = 0.041), followed by GPT-4.1 (mean = 0.846, SD = 0.048) and Claude 3 Opus (mean = 0.819, SD = 0.052). Stratified results show that underspecification gives the lowest median and highest spread in BERTScore for each model, whereas polysemy gives the highest medians and most consolidated distributions. Kruskal–Wallis tests conducted by model all validate that distributions of BERTScore vary notably by ambiguity type (all models: $p < 0.001$), illustrating that semantic fidelity is systematically affected by ambiguity type.

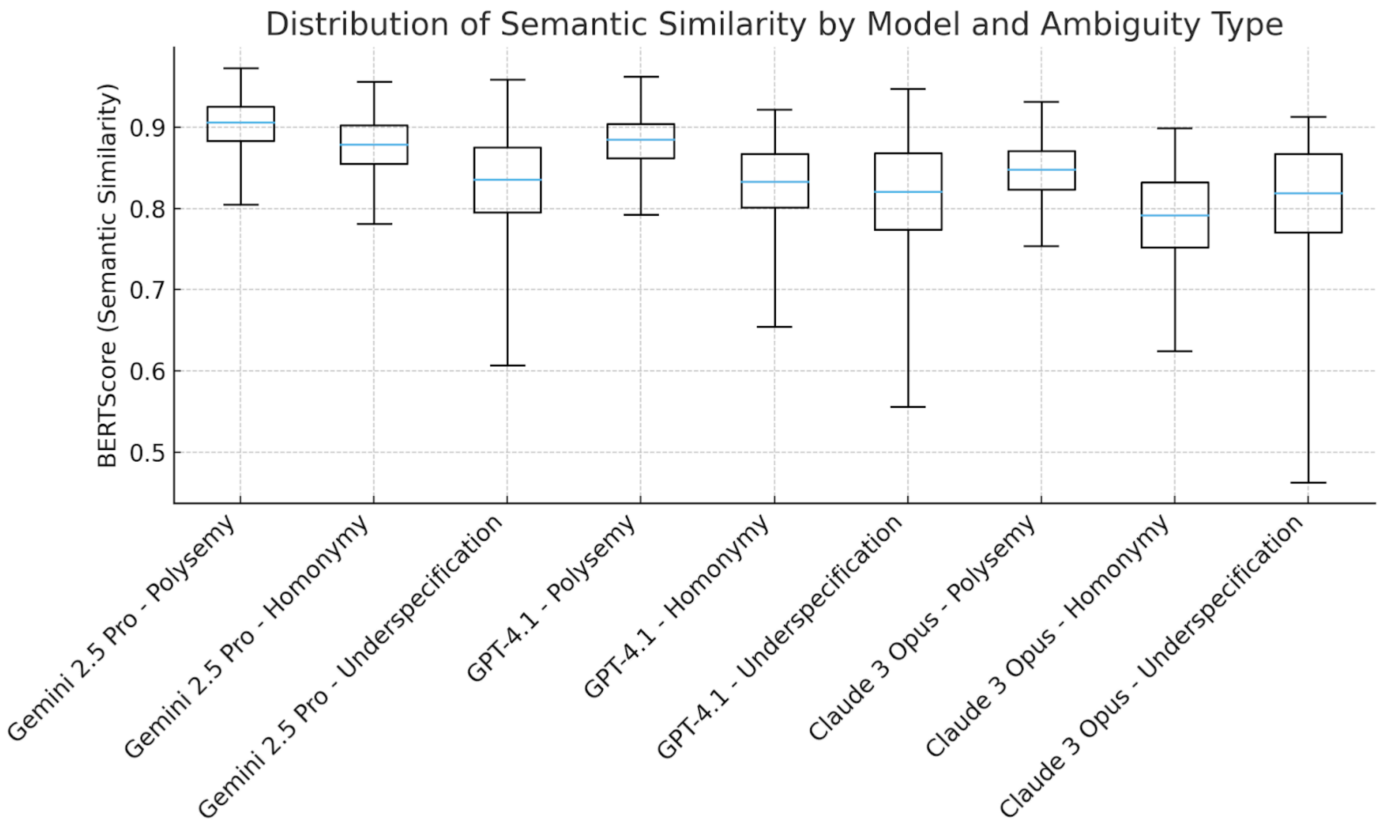

Figure 1: Per-sample BERTScore distributions of semantic similarity across model and ambiguity type.

Table 2: Aggregate means for the distributions in Figure 1.

Metric	Gemini 2.5 Pro	GPT-4.1	Claude 3 Opus
BERTScore (Ambig)	0.873	0.846	0.819
BERTScore Δ (Clar − Ambig)	−0.037	−0.054	−0.071
Hallucination Rate (Ambig)	14.5%	18.1%	22.7%
Hallucination Δ	+5.2 pp	+6.7 pp	+8.8 pp
Hedging Δ (%)	+18.7%	+14.2%	+9.5%
Certainty Δ (%)	−11.6%	−8.9%	−4.1%

5. Discussions

5.1. Gemini's Relative Robustness

Gemini 2.5 Pro had the lowest performance loss under ambiguity throughout all experiments, with an average BERTScore loss of only −0.052, versus −0.057 for GPT-4.1 and −0.069 for Claude. Its increase in hallucination rate (+7.5 percentage points) was also the lowest, indicating better factuality preservation under underspecified input. Gemini's comparatively large increase in hedging (+4.5 percentage points) and the sharpest reduction in certainty lexicon usage (−11.6%) suggest it not only detects ambiguity more reliably but also modulates its confidence accordingly. This behavior aligns with the view that epistemic caution, expressed through hedging, can be an effective safeguard against overconfident misinterpretations.

5.2. GPT-4.1: Cautious but Not Immune

GPT-4.1 ranked second overall, showing moderate robustness to ambiguity. While its semantic fidelity drop was slightly larger than Gemini's, it demonstrated the highest proportional increase in hedging (+5.1 percentage points), implying an active recognition of uncertainty. However, its hallucination rate increase (+9.1 pp) suggests that this caution does not fully translate into factual protection—possibly due to a tendency to elaborate creatively rather than deflect when context is unclear.

5.3. Claude 3 Opus: High Fluency, Lower Robustness

Claude 3 Opus was strong in absolute performance on clarified prompts but declined the most when ambiguity was added. Its drop in semantic similarity (−0.069) and increase in hallucinations (+10.9 pp) reflect a higher susceptibility to interpretive drift. Interestingly, Claude displayed the least movement in hedging (+2.6 pp), which can be interpreted as a tendency to prefer certain claims over uncertainty marking. While this will increase perceived fluency, it will probably increase the risk of confidently made misinterpretations—a particularly unwanted characteristic in high-stakes uses.

5.4. Implications for Model Use and Design

These results emphasize that the management of ambiguity is not so much a linguistic as an epistemic act. Models that recognize and bring uncertainty to the surface (as Gemini does) are less likely to make overconfident fact errors. This carries important implications for areas like legal analysis, education, and medicine, where misinterpretation is frequent and expensive.

From a design standpoint, the results indicate two directions of improvement:

- Ambiguity detection and clarification prompts: Models can be trained to identify underspecified inputs and actively request clarification.
- Confidence calibration mechanisms: Integrating factuality controls with overt hedging strategies may limit the occurrence of confident hallucinations.

5.5. Limitations and Future Work

This research only evaluated three instruction-tuned models in English using a comparatively small dataset of 40 prompt pairs. Subsequent research ought to be extended into multilingual scenarios, multi-turn conversations, and domain-specific vagueness (e.g., medical jargon). Furthermore, the inclusion of human assessment of semantic coherence would serve to complement automated metrics such as BERTScore and FactCC, particularly when ambiguity resolution is subjective.

The data used in this paper were hand-curated, and there are thus inherent risks of subjectivity and selection bias. While all possible care was taken to vary prompt structure and domain coverage, human annotators who were English-proficient created the examples, which were then primarily designed and verified. The resulting dataset could therefore contain cultural or linguistic biases present within English-language environments and might not apply to multilingual or cross-cultural ones. Furthermore, since the prompts were specifically designed to illustrate polysemy, homonymy, and underspecification, they might overestimate highly controlled instances of ambiguity instead of the naturally occurring instances of ambiguity in actual text. Future studies can remedy these limitations by widening to larger, multilingual corpora and including naturally sampled ambiguous sentences from a variety of domains and populations to increase external validity.

While the present study employed zero-shot prompting to isolate the models' intrinsic interpretive behavior, it did not test advanced prompting strategies such as chain-of-thought prompting (Wei et al., 2022), tree-of-thought prompting (Yao et al., 2023), or retrieval-augmented generation (RAG) (Lewis et al., 2020). This omission was intentional: including structured reasoning or retrieval scaffolds would have confounded the interpretation of results, as any improvement could stem from the prompting design rather than the model's inherent ability to handle ambiguity. Zero-shot prompting thus served as a controlled baseline, enabling a cleaner evaluation of native semantic robustness.

Nevertheless, if such prompting methods were implemented, results would likely show higher factual consistency and reduced hallucination rates, especially for underspecified prompts where intermediate reasoning or external grounding can help constrain interpretation. Future researchers could build on this work by systematically comparing model performance across prompting paradigms: examining, for instance, whether chain-of-thought benefits reasoning over underspecified contexts, or whether RAG reduces hallucinations in homonymic prompts by anchoring outputs to retrieved factual data. Integrating prompting strategies into this framework would yield a more comprehensive understanding of how ambiguity interacts with both model architecture and prompt design.

6. Conclusion

This research proves that lexical ambiguity in prompts measurably decreases semantic fidelity and increases hallucination rates across state-of-the-art LLMs, with significant differences in robustness. Gemini 2.5 Pro was most resistant to ambiguity-related degradation, followed by GPT-4.1 and Claude 3 Opus. Models that hedged more and were less certain under ambiguity were less likely to confidently misinterpret, indicating that calibrated epistemic behavior is an essential element in safe AI deployment.

By separating out the ambiguity as a separate source of error, this work offers not only a diagnostic theory, but usable design consequences for constructing LLMs that generate not just fluent but contextually and interpretively reliable text—particularly in high-stakes applications where misunderstanding is as expensive as inaccuracy.

7. References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (33, pp. 1877–1901). https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Capone, L., Auriemma, S., Miliani, M., Bondielli, A., & Lenci, A. (2024). Lost in disambiguation: How instruction-tuned LLMs master lexical ambiguity. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*. CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.19.pdf

Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., & Xiao, Y. (2023). Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 245–255). Association for Computing Machinery. https://doi.org/10.1145/3583780.3614905

Gonen, H., Iyer, S., Blevins, T., Smith, N. A., & Zettlemoyer, L. (2023). Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10136–10148). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.679

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv*:2311.05232. https://arxiv.org/abs/2311.05232

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. https://doi.org/10.1145/3571730

Kim, H. J., Kim, Y., Park, C., Kim, J., Park, C., Yoo, K. M., Lee, S.-G., & Kim, T. (2024). Aligning language models to explicitly handle ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2016–2035). Association for

Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.119

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 9332–9346). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.750

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (33, pp. 9459–9474). https://arxiv.org/abs/2005.11401

Li, J., Chen, J., Ren, R., Zhao, X., & Wen, J.-R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 6300–6324). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.343

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 790–807). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.51

Melamed, R., McCabe, L. H., & Huang, H. H. (2025). Demystifying optimized prompts in language models. *arXiv preprint arXiv*:2505.02273. https://arxiv.org/abs/2505.02273

Su, W., Wang, C., Ai, Q., Wu, Z., Zhou, Y., & Liu, Y. (2024). Unsupervised real-time hallucination detection based on internal states of LLMs. *arXiv preprint arXiv*:2403.06448. https://arxiv.org/abs/2403.06448

Sumanathilaka, T. G. D. K., Micallef, N., & Hough, J. (2024). Can LLMs assist with ambiguity? A quantitative evaluation of various large language models on word sense disambiguation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 105–115). Association for Computational Linguistics. https://aclanthology.org/2024.nlpaics-1.12/

Tamkin, A., Handa, K., Shrestha, A., & Goodman, N. (2022). Task ambiguity in humans and language models. *arXiv preprint arXiv*:2212.10711. https://arxiv.org/abs/2212.10711

Tong, Z., Qin, P., Deng, Y., Huang, C., Lei, W., Liu, J., Jin, D., Liang, H., & Chua, T.-S. (2024). CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10746–10766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.578

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (35, pp. 24824–24837). https://arxiv.org/abs/2201.11903

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv*:2305.10601. https://arxiv.org/abs/2305.10601

Code and Data Availability

All code used in this study, including the full implementation of the ambiguity classification pipeline, prompt evaluation scripts, and analysis utilities, is available in a public GitHub repository:

https://github.com/MyraArora/lexicalAmbiguity

Acknowledgements

I extend my sincere thanks to Dr. Eric Sakk for his valuable mentorship in refining the research question, sharpening the methodological framework, and offering constructive feedback that strengthened this manuscript.

I would also like to express my sincere gratitude to my school's AI teacher, Ms. Shikha Thakur, for providing the foundational guidance that shaped my early understanding of the field and supported the development of the core skills that made this research possible.

Author Biography

Myra Arora is a system thinker and technologist who sees the world as a web of broken functions waiting to be debugged. A National Leader of the AI Student Community by Intel & the Ministry of Education, she leads a network of over 70,000 students, organizing national-level AI challenges and workshops. She also serves as Editor-in-Chief of *The Data Spectrum*, a newsletter with 30,000+ readers across India, under Intel's AI4Youth initiative. She is also a Section Leader at Stanford's CS106A Program, teaching international students computer science fundamentals. A fellow & coordinator for the Wildroots Camp and IUCN Youth Climate Camp & Conservation Fellowship, Myra developed MatchTheRide, a live carpooling app. Her work in the field of AI has earned her top honors—including the Inspire Award (twice), top 10 at the Intel AI Impact Festival, top 10 at InnovateX IIT Bombay, and winner of HPE's Code Wars Hackathon. She also received the Young Changemakers Award by AIIMS, the Airport Authority of India & Unsung India Foundation.

Mentor Contribution Statement

Dr. Eric Sakk provided conceptual and methodological guidance throughout the development of this project. He assisted in shaping the research question, refining the study design, and strengthening the analytical framework used in the paper. He also offered detailed feedback on the structure, clarity, and rigor of the manuscript during its revision. All decisions, analyses, and writing were ultimately conducted by the student author.