Prompt Sensitivity of Hallucinations in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) frequently produce hallucinations—factually incorrect outputs that appear fluent and confident. While prior work focuses on model and data factors, the role of prompt formulation remains underexplored. We identify **prompt-induced hallucination instability**, where small changes in prompt phrasing lead to different factual outcomes for the same query.

We introduce a controlled evaluation framework using diverse prompt templates and likelihood-based scoring on TruthfulQA and HaluEval. We find that factual correctness is highly unstable, with **19%–60% of examples flipping** between correct and incorrect responses across prompts. These effects are strongly model- and task-dependent, challenging the reliability of prompt engineering as a mitigation strategy.

Extending to free-form generation with an LLM-based judge, we show that instability **persists and is often amplified**, with flip rates reaching **32%–45% on HaluEval and 61%–69% on TruthfulQA**. Finally, we provide a mechanistic explanation linking instability to decision margins, showing that prompt-induced flips concentrate near cases where models assign similar likelihood to competing answers. Our findings demonstrate that hallucination is not a fixed model property, but an emergent behavior shaped by prompt–model interactions.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of natural language tasks, but they remain prone to *hallucinations*—outputs that are fluent and plausible yet factually incorrect or unfaithful. This issue has been widely documented, with benchmarks such as TruthfulQA showing that even state-of-the-art models can confidently generate false statements (Lin et al., 2022). More recent efforts, including large-scale evaluation frameworks such as HaluEval, further highlight the prevalence of hallucinations across tasks and settings (Li et al., 2023a). In response, prior work has focused on detecting hallucinations using consistency checks (Manakul et al., 2023), uncertainty-based methods (Farquhar et al., 2024), and model-internal signals (Sriramanan et al., 2024).

While these approaches primarily treat hallucination as a property of the model or its training data, the role of *prompt formulation* remains less well understood. Prompts are the primary interface between users and LLMs, and small changes in phrasing can lead to different outputs. Prior work on prompt sensitivity has examined variability in outputs or task accuracy (Zhuo et al., 2024), but does not directly address variability in *factual correctness*. In the context of hallucinations, the key concern is not whether outputs differ, but whether they remain factually reliable under small prompt changes. This raises a critical question: can the same model produce both correct and hallucinated answers for the same query solely due to prompt variation?

In this work, we investigate **prompt-induced hallucination instability**—the phenomenon where different prompts for the same underlying query lead to different factual outcomes. We propose a controlled evaluation framework that isolates prompt effects by holding the query and answer space fixed while varying prompt formulations. To enable consistent comparison, we adopt a likelihood-based protocol that measures model preference between correct and hallucinated answers, and introduce metrics such as flip rate and Hallucination Instability Score (HIS) to quantify instability.

We evaluate instruction-tuned LLMs across two complementary benchmarks: TruthfulQA, which emphasizes knowledge-intensive generation, and HaluEval, which focuses on hallucination detection

in a multiple-choice setting. We find that hallucination outcomes are highly sensitive to prompt variation, with up to **60%** of examples flipping between correct and incorrect responses across prompts. These effects are both *model-dependent* and *task-dependent*, with no consistent prompt strategy that reliably improves factual accuracy.

To assess whether these effects persist in realistic settings, we conduct a free-form generation study evaluated using an LLM-based judge. We find that instability not only persists but often increases under free-form generation, indicating that prompt sensitivity is not an artifact of controlled evaluation. Finally, we provide a mechanistic analysis showing that instability is concentrated near decision boundaries, where small likelihood differences between correct and incorrect answers make model predictions sensitive to prompt perturbations.

These findings suggest that hallucination is not a stable property of LLMs, but an outcome of interactions between prompts, models, and tasks. More broadly, they reveal a failure mode that is not captured by standard robustness evaluations, which may report stable accuracy while masking variability in factual correctness. Our work highlights the need to treat prompt sensitivity as a first-class consideration in evaluating and mitigating hallucinations in LLMs.

2 Related Work

2.1 Hallucination in Large Language Models

Hallucinations—outputs that are fluent but factually incorrect—remain a central challenge in large language models (LLMs). Early work has shown that even strong models can confidently produce false statements, often reflecting patterns in training data rather than factual correctness (Lin et al., 2022). Subsequent benchmarks such as HaluEval extend this line of work by providing large-scale evaluation settings that systematically probe hallucination behavior across tasks (Li et al., 2023a). More recent efforts, including HalluLens and multilingual shared tasks such as Mu-SHROOM, further highlight the diversity of hallucination phenomena across domains, languages, and evaluation setups (Bang et al., 2025; Vazquez et al., 2025). These works establish hallucination as a persistent and multifaceted problem, motivating the need for robust evaluation and mitigation strategies.

2.2 Hallucination Detection and Uncertainty Estimation

A significant body of work has focused on detecting hallucinations without relying on external knowledge sources. SelfCheckGPT proposes a black-box approach based on response consistency under stochastic sampling, showing that hallucinated outputs tend to exhibit higher variability (Manakul et al., 2023). Other approaches leverage uncertainty signals, such as semantic entropy, to identify unreliable generations at the semantic level (Farquhar et al., 2024). Follow-up work introduces lightweight variants, including semantic entropy probes, which approximate these signals more efficiently (Kossen et al., 2024). Additional methods exploit internal model representations or token-level signals for hallucination detection (Zhang et al., 2024; Niu et al., 2025). Recent approaches such as LLM-Check further demonstrate that even single-response signals can provide meaningful hallucination detection performance (Sriramanan et al., 2024). While effective, these methods primarily treat hallucination as a property of model outputs, rather than considering variability induced by prompt formulation.

2.3 Prompt Sensitivity and Evaluation Robustness

Prompt formulation has been shown to significantly influence LLM behavior across tasks. Prior work has demonstrated that model outputs can vary substantially under different prompt paraphrases, raising concerns about evaluation robustness (Zhuo et al., 2024). More recent studies argue that some observed prompt sensitivity may stem from evaluation artifacts rather than intrinsic model behavior (Hua et al., 2025). In our prior work (Authors, 2026), we observed that prompt phrasing can affect hallucination outcomes, suggesting that hallucination behavior may not be stable across prompts. However, existing work has largely focused on performance variability or evaluation consistency, rather than systematically quantifying how prompt variation affects factual correctness across models and tasks.

2.4 Mitigation Strategies for Hallucination

Various approaches have been proposed to mitigate hallucinations in LLMs. These include prompt engineering techniques, such as structured prompting and multi-stage prompt refinement, which

aim to guide models toward more reliable outputs (Shim et al., 2025). Other methods focus on training-time interventions, such as sensitivity dropout, to improve robustness to perturbations in inputs (Mohammadzadeh et al., 2025). Additionally, reasoning-based prompting and external knowledge augmentation have been explored as ways to reduce hallucination in specific settings (Li et al., 2023a). Despite these efforts, existing mitigation strategies often assume that prompt-based interventions generalize across models and tasks.

2.5 Our Contribution

We introduce *prompt-induced hallucination instability* as a measurable property of large language models. Unlike prior work on prompt sensitivity, which focuses on output variation or task accuracy, we study *variation in factual correctness* under prompt changes.

We propose a controlled evaluation framework that isolates prompt effects and define quantitative metrics such as flip rate and Hallucination Instability Score (HIS) to measure instability across prompts, models, and datasets.

We further show that this instability persists—and often increases—under free-form generation, indicating that it is not an artifact of controlled evaluation. Finally, we provide a simple mechanistic analysis linking instability to decision margins, showing that flips are concentrated near uncertain predictions.

Together, our work moves from describing prompt sensitivity to quantitatively measuring and explaining factual instability under prompt perturbations.

3 Methodology

3.1 Overview

Our goal is to measure *prompt-induced hallucination instability*, defined as the variability in factual correctness of model outputs under different prompt formulations for the same underlying query. To this end, we design a controlled evaluation framework that systematically varies prompt templates while holding the query and answer space fixed.

We first evaluate model behavior using a likelihood-based scoring protocol, enabling consistent comparison across prompts without relying on free-form generation. This controlled setup allows us to isolate the effect of prompt formulation on model preference between correct and hallucinated answers.

To assess whether these effects generalize beyond controlled settings, we additionally conduct a free-form evaluation, where models generate answers directly and outputs are evaluated using an LLM-based judge. This complementary setup enables us to validate whether prompt-induced instability persists under more realistic usage conditions.

3.2 Prompt Template Design

We construct a set of prompt templates spanning multiple families that reflect common prompting strategies in LLM usage:

- **Neutral**: direct question-answer format without additional guidance.
- **Reasoning**: prompts that encourage step-by-step reasoning or intermediate explanations.
- **Uncertainty-aware**: prompts that explicitly allow or encourage the model to express uncertainty.
- **Structured**: prompts that enforce constrained output formats, such as JSON-style responses.

Each template is instantiated across all examples, resulting in multiple prompt variants per query. This design allows us to isolate the effect of prompt formulation on hallucination outcomes. Our prompt families are inspired by widely used prompting strategies such as chain-of-thought reasoning (Wei et al., 2022) and structured output prompting in LLM applications. Full prompt templates are provided in Appendix A.

Prompt Strength. We define prompt strength as the degree of guidance or constraint imposed by the prompt on the model's response. We assign each template a discrete strength level from 1 to 4, where strength 1 corresponds to minimal instruction (e.g., direct question answering), and strength 4 corresponds to highly constrained formats (e.g., structured JSON outputs). Intermediate levels include light formatting constraints (e.g., one-sentence answers) and behavioral guidance such as step-by-step reasoning or confidence specification. This allows us to study whether increasing prompt constraint leads to more stable or more accurate outputs.

3.3 Evaluation Protocol

To ensure consistent and comparable evaluation across prompts, we adopt a **multiple-choice likelihood-based scoring framework**. For each query, we construct two candidate answers:

- A **correct** answer consistent with ground truth.
- A **hallucinated** (incorrect) answer.

Given a prompt p and candidate answers a_{correct} and $a_{\text{hallucinated}}$, we compute the conditional log-likelihood of each answer under the model:

$$\log P(a \mid p) \tag{1}$$

The model's prediction is defined as the answer with higher likelihood. This approach avoids reliance on free-form generation and ensures deterministic evaluation, isolating prompt effects from decoding variability (Li et al., 2023a). As such, it provides a controlled setting for analyzing prompt-induced differences in model behavior.

We define a binary hallucination indicator for each prompt-instance:

$$h = \mathbf{1}\left[\log P(a_{\text{hallucinated}} \mid p) > \log P(a_{\text{correct}} \mid p)\right] \tag{2}$$

where $h = 1$ indicates a hallucinated outcome. While this formulation enables precise comparison across prompts, it does not capture the full complexity of free-form generation. We therefore complement this analysis with a free-form evaluation using an LLM-based judge, described later.

3.4 Free-form Evaluation with LLM Judge

To validate whether prompt-induced instability extends beyond controlled likelihood-based evaluation, we conduct a complementary free-form generation study. In this setting, models are prompted to generate answers directly, without constraining outputs to a fixed candidate set. Due to generation cost constraints, free-form evaluation is conducted on three representative models (Qwen-2.5-7B, Phi-3-mini, and LLaMA-3.1-8B) spanning different performance levels; trends are consistent with the controlled setting.

For each query, we apply the same set of prompt templates used in the likelihood-based evaluation, ensuring consistency in prompt variation across settings. The resulting model outputs are then evaluated for factual correctness using an LLM-based judge. Specifically, we provide the judge with the question, a reference (ground-truth) answer, and the model-generated response, and ask it to determine whether the response is *correct*, *incorrect*, or *unknown*. We use a GPT-based model (GPT-4o-mini) as the evaluator in this setting.

This evaluation protocol follows recent work on LLM-as-a-judge frameworks (Zheng et al., 2023; Li et al., 2023b), which have been shown to provide reliable assessments of factual correctness and response quality in free-form outputs. We use deterministic decoding (temperature=0) and structured prompts to ensure consistent judgments.

We convert the judge outputs into binary hallucination indicators, treating *incorrect* responses as hallucinations and excluding *unknown* cases from certain analyses. Using these labels, we compute the same instability metrics as in the likelihood-based setting, enabling direct comparison between controlled and free-form evaluations. Additional details are provided in Appendix B.

3.5 Instability Metrics

To quantify prompt-induced variability, we introduce several metrics computed across prompt variants for each query.

Hallucination Rate. We compute the mean hallucination rate across all prompt-instance pairs.

Flip Rate. We define the flip rate as the fraction of queries for which hallucination outcomes differ across prompt templates:

$$\text{FlipRate} = \mathbb{E}_x\left[\mathbf{1}\left(\exists i, j : h_{x,i} \neq h_{x,j}\right)\right] \tag{3}$$

This measures how often prompt variation changes the factual correctness of model outputs. We treat incorrect responses as hallucinations and compute flips over binary correctness labels; cases labeled as unknown are rare and do not materially affect results.

Hallucination Instability Score (HIS). We define the Hallucination Instability Score as:

$$\text{HIS}(x) = 2p_x(1 - p_x) \tag{4}$$

where p_x is the fraction of prompt variants for query x that result in hallucination. This quantity is maximized when outcomes are evenly split, capturing maximum instability.

Prompt Gap. For each query, we compute the difference between the worst and best prompt outcomes:

$$\text{Gap}(x) = \max_i h_{x,i} - \min_i h_{x,i} \tag{5}$$

A non-zero gap indicates that at least one prompt yields a correct answer while another yields a hallucination.

3.6 Datasets and Models

We evaluate our framework on two complementary benchmarks:

- **TruthfulQA** (Lin et al., 2022), which focuses on knowledge-intensive questions prone to common misconceptions.
- **HaluEval** (Li et al., 2023a), a large-scale benchmark designed for hallucination detection in a multiple-choice setting.

We conduct experiments on five instruction-tuned LLMs spanning multiple model families, including LLaMA-based models (Touvron et al., 2023), Qwen models (Jiang et al., 2023), and other recent open-source instruction-tuned architectures. This diversity allows us to study both model-specific and cross-model effects of prompt-induced hallucination instability.

4 Results

4.1 Controlled Instability under Likelihood-based Evaluation

We first evaluate prompt-induced hallucination instability under a controlled likelihood-based framework. Table 1 summarizes hallucination rates, flip rates, and instability scores across models and datasets.

We observe that hallucination outcomes are highly sensitive to prompt formulation. Across models, between **19% and 60%** of examples flip between correct and incorrect predictions under different prompt variants. This demonstrates that factual correctness is not a stable property of the model for a given query, but depends strongly on prompt phrasing.

Instability is consistently higher on TruthfulQA compared to HaluEval, indicating that prompt sensitivity is amplified for knowledge-intensive and adversarial queries. Notably, some models achieve relatively low hallucination rates while still exhibiting substantial flip rates, suggesting that accuracy alone does not imply robustness to prompt variation.

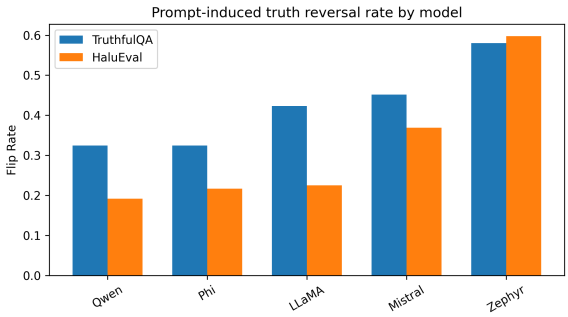


Figure 1: Flip rate across models, showing the fraction of examples whose correctness changes across prompts.

As shown in Figure 1, a substantial fraction of examples exhibit changes in correctness across prompt formulations.

We report bootstrap 95% confidence intervals for flip rates across all models and datasets. Intervals are consistently narrow (within $\pm 1\%$ on HaluEval and ± 3–4% on TruthfulQA), indicating that the observed differences are statistically robust (see Appendix C for details).

4.2 Effect of Prompt Design

We next analyze how prompt design influences hallucination behavior across models. Table 2 presents hallucination rates grouped by prompt family, and Table 3 provides a more fine-grained view at the level of individual templates.

We find that prompt effects are highly variable across models and do not follow a consistent pattern. Structured prompts, such as JSON-based outputs, often increase hallucination rates, particularly on knowledge-intensive tasks. Similarly, reasoning-based prompts do not consistently improve factual correctness and can, in some cases, introduce additional errors. Uncertainty-aware prompts show mixed behavior, improving performance for some models while degrading it for others.

Importantly, increasing prompt complexity or constraint does not reliably lead to better performance. Across models, simpler prompts frequently perform as well as or better than more structured or guided alternatives. These findings suggest that prompt effectiveness depends on complex interactions between model architecture and task characteristics, rather than following a universal pattern.

These observations raise an important question: are these effects an artifact of the controlled evalu-

Dataset	Model	Hallucination ↓	Flip Rate ↑	HIS ↑	Gap ↑	Any Gap ↑
HaluEval	Qwen2.5-7B	0.136	0.192	0.072	0.192	0.192
	LLaMA-3.1-8B	0.138	0.225	0.085	0.225	0.225
	Phi-3-mini	0.145	0.217	0.084	0.217	0.217
	Mistral-7B	0.219	0.369	0.141	0.369	0.369
	Zephyr-7B	0.280	0.598	0.221	0.598	0.598
TruthfulQA	Qwen2.5-7B	0.285	0.324	0.122	0.324	0.324
	LLaMA-3.1-8B	0.352	0.424	0.159	0.424	0.424
	Phi-3-mini	0.380	0.324	0.120	0.324	0.324
	Mistral-7B	0.371	0.452	0.170	0.452	0.452
	Zephyr-7B	0.396	0.580	0.221	0.580	0.580

Table 1: Hallucination and instability metrics across models and datasets. Hallucination rate measures the fraction of incorrect outcomes. Flip rate measures the fraction of examples whose correctness changes across prompts. HIS (Hallucination Instability Score) captures variability across prompts.

Dataset	Model	F	R	B	U
HaluEval	Qwen	0.148	0.123	0.123	0.136
	LLaMA	0.127	0.117	0.141	0.158
	Phi	0.155	0.134	0.144	0.140
	Mistral	0.235	0.213	0.235	0.197
	Zephyr	0.334	0.267	0.255	0.243
TruthfulQA	Qwen	0.291	0.259	0.252	0.309
	LLaMA	0.399	0.350	0.326	0.320
	Phi	0.440	0.307	0.364	0.365
	Mistral	0.434	0.355	0.329	0.338
	Zephyr	0.461	0.389	0.371	0.347

Table 2: Hallucination rate across prompt families (F=Format, R=Reasoning, B=Base, U=Uncertainty). Lower is better.

ation setup, or do they persist under realistic generation conditions? We address this next. Figure 2

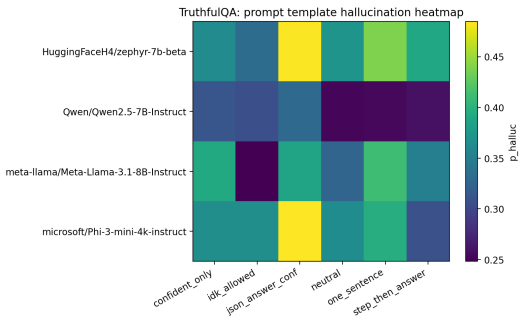


Figure 2: Hallucination rates across prompt templates on TruthfulQA. Performance varies significantly across prompt formulations, highlighting strong prompt sensitivity and the absence of a consistently reliable prompt design.

illustrates this variability, showing that no single prompt consistently minimizes hallucination across models.

4.3 Free-form Validation with LLM Judge

To assess the robustness of our findings, we conduct a free-form evaluation where models generate answers directly and outputs are evaluated using a GPT-based judge. Table 4 presents the results.

We find that prompt-induced instability remains **consistently high under free-form generation**, closely mirroring the trends observed in the likelihood-based setting. This demonstrates that our findings are not an artifact of the evaluation protocol, but reflect a fundamental property of model behavior.

On HaluEval, hallucination rates remain relatively low (approximately 14–15%), yet flip rates range from **32% to 45%**. This indicates that even when models are generally accurate, prompt variation frequently changes the correctness of outputs.

On TruthfulQA, both hallucination and instability increase substantially. Flip rates reach **61% to 69%**, demonstrating that prompt sensitivity becomes a dominant factor for more challenging queries.

To assess the reliability of the LLM-based judge, we manually annotated a random sample of 40 examples and observed an agreement rate of 0.95 with the GPT-based labels (Cohen's $\kappa = 0.89$), indicating strong consistency.

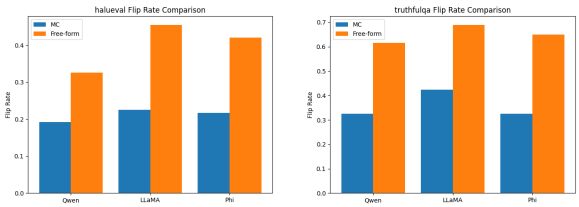
Consistency across evaluation settings. We observe that models exhibiting higher instability under likelihood-based evaluation also tend to exhibit higher instability under free-form generation. For example, LLaMA-3.1-8B shows consistently higher flip rates than Qwen across both settings. This alignment suggests that prompt-induced instability is an intrinsic model property rather than an artifact of a particular evaluation protocol.

Model	Family	Template	Strength	p_truth ↑	p_halluc ↓
Qwen	Base	Neutral	1	0.748	0.252
Qwen	Format	One Sentence	2	0.747	0.253
Qwen	Reasoning	Step-then-Answer	3	0.741	0.259
Qwen	Uncertainty	IDK Allowed	2	0.694	0.306
Qwen	Uncertainty	Confident Only	3	0.688	0.312
Qwen	Format	JSON	4	0.671	0.329
LLaMA	Base	Neutral	1	0.674	0.326
LLaMA	Uncertainty	IDK Allowed	2	0.752	0.248
LLaMA	Reasoning	Step-then-Answer	3	0.650	0.350
LLaMA	Format	JSON	4	0.613	0.387
LLaMA	Uncertainty	Confident Only	3	0.608	0.392
LLaMA	Format	One Sentence	2	0.589	0.411

Table 3: Effect of individual prompt templates and strength levels on hallucination behavior (TruthfulQA).

Dataset	Model	Halluc. ↓	Flip ↑	HIS ↑	Unknown ↓
HaluEval	Qwen-2.5-7B	0.143	0.325	0.117	0.000
	Phi-3-mini	0.148	0.420	0.146	0.005
	LLaMA-3.1-8B	0.153	0.455	0.156	0.000
TruthfulQA	Qwen-2.5-7B	0.487	0.615	0.238	0.012
	Phi-3-mini	0.606	0.650	0.258	0.074
	LLaMA-3.1-8B	0.691	0.690	0.259	0.046

Table 4: Free-form evaluation results using a GPT-based judge. We report hallucination rate, flip rate, Hallucination Instability Score (HIS), and the fraction of outputs labeled as unknown. Prompt-induced instability persists across models and datasets, with especially high flip rates on TruthfulQA.



(a) HaluEval (b) TruthfulQA

Figure 3: Comparison of flip rates across likelihood-based (MC) and free-form evaluation.

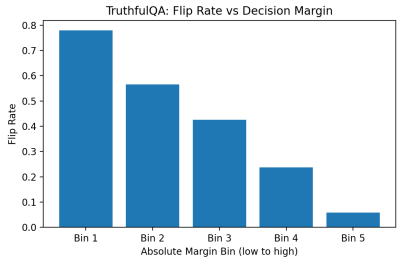


Figure 4: Flip rate as a function of decision margin on TruthfulQA. Flip rates decrease monotonically as margin increases, indicating that prompt-induced instability is concentrated near decision boundaries.

4.4 Instability and Decision Margin

To better understand why certain examples are more likely to exhibit prompt-induced instability, we analyze the relationship between instability and the likelihood margin between correct and hallucinated answers. Figure 4 shows that examples with smaller margins are substantially more likely to flip across prompt variants, whereas examples with larger margins remain stable. In particular, the lowest margin bin exhibits a flip rate of over 75%, while the highest margin bin shows near-zero instability.

This trend suggests that prompt-induced instability is concentrated near decision boundaries, where the model assigns similar likelihood to correct and incorrect answers. In such cases, small changes in prompt formulation can shift the model's preference between competing responses.

These findings provide a mechanistic explanation for prompt sensitivity, linking instability to underlying model uncertainty rather than arbitrary prompt effects.

5 Discussion

Our results demonstrate that hallucination behavior in large language models is not a fixed property of the model alone, but emerges from interactions between the model, prompt formulation, and task setting. Across both likelihood-based and free-form

evaluations, we observe substantial variability in factual correctness under different prompt formulations, indicating that prompt-induced instability is a pervasive phenomenon.

Importantly, this instability persists—and often increases—under free-form generation. Flip rates are consistently higher in the free-form setting across all models and datasets, suggesting that prompt sensitivity is amplified when models generate unconstrained outputs. This has direct implications for real-world usage, where models are typically deployed in open-ended settings rather than controlled evaluation environments.

We also find that prompt design does not follow a consistent pattern across models. Strategies such as structured formatting or step-by-step reasoning do not reliably improve factual accuracy and can, in some cases, introduce additional errors. This challenges the common assumption that more complex or guided prompting leads to more reliable outputs, and suggests that prompt effectiveness depends on nuanced interactions with model behavior.

From an evaluation perspective, our findings highlight the limitations of relying on a single prompt or evaluation setup. Models that appear reliable under one prompt may exhibit significantly different behavior under another, indicating that single-prompt evaluations may provide an incomplete view of model reliability.

Implications for mitigation. Our findings suggest that mitigating hallucinations requires approaches that are robust to prompt variation, rather than relying on a single carefully engineered prompt. One practical direction is to use *prompt ensembling*, where multiple prompt formulations are evaluated and aggregated to reduce sensitivity to any single prompt. Another direction is to incorporate *uncertainty-aware mechanisms*, such as abstention or confidence estimation, particularly for low-margin cases that are more likely to exhibit instability. More broadly, our results highlight the importance of developing *prompt-invariant evaluation protocols*, where model reliability is assessed across diverse prompt formulations rather than a fixed template.

Overall, our results underscore the need to treat prompt sensitivity as a first-class concern in both evaluation and deployment. Robust assessment of LLM reliability should account for variability across prompt formulations, rather than relying on single-prompt measurements.

6 Limitations

Our study has several limitations. First, we evaluate a fixed set of prompt templates, which, while representative, do not cover the full space of possible prompt variations. Second, our categorization of prompt families is heuristic and may not capture finer-grained differences in prompt structure. Third, our free-form evaluation relies on an LLM-based judge (GPT-4o-mini), which may introduce variability for borderline cases, although aggregate trends remain consistent. Finally, we consider a limited set of open-weight instruction-tuned models, and results may not fully generalize to larger proprietary systems or differently aligned models.

7 Conclusion

We introduced the concept of *prompt-induced hallucination instability* and presented a systematic framework for evaluating how prompt formulation affects factual correctness in large language models. Across multiple models and datasets, we showed that hallucination outcomes vary substantially across prompts, with a large fraction of examples changing correctness under different prompt formulations.

Crucially, we demonstrated that this instability is not limited to controlled evaluation settings, but persists—and often increases—under free-form generation. This indicates that prompt sensitivity is a fundamental characteristic of model behavior rather than an artifact of evaluation methodology.

Our results further show that prompt effects are both model-dependent and task-dependent, and that commonly used prompting strategies, including structured and reasoning-based prompts, do not reliably improve factual accuracy. These findings challenge the assumption that prompt engineering alone can serve as a robust solution for mitigating hallucinations.

Overall, our work highlights the need to treat prompt sensitivity as a first-class consideration in both evaluation and deployment of LLMs. Future work should focus on developing evaluation protocols and modeling approaches that explicitly account for—and reduce—instability under prompt variation.

References

Anonymous Authors. 2026. Context-length robustness in question answering models: A comparative empir-

ical study. Accepted at Math AI 2026.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. 2025. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19889–19899, Suzhou, China. Association for Computational Linguistics.

Albert Q. Jiang and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *Preprint*, arXiv:2406.15927.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*.

Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, and Golnoosh Farnadi. 2025. Hallucination detox: Sensitivity dropout (SenD) for large language model training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5538–5554, Vienna, Austria. Association for Computational Linguistics.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. 2025. Robust hallucination detection in llms via adaptive token selection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jung-Woo Shim, Yeong-Joon Ju, Ji-Hoon Park, and Seong-Whan Lee. 2025. Multi-stage prompt refinement for mitigating hallucinations in large language models. *arXiv preprint arXiv:2510.12032*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. In *arXiv preprint arXiv:2302.13971*.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: MuSHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2024. Prompt-guided internal states for hallucination detection of large language models. *arXiv preprint arXiv:2411.04847*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages

1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt Templates

We provide the full set of prompt templates used in our experiments. Each template is applied to the same underlying query to isolate the effect of prompt formulation. Templates are grouped by prompt family and vary in strength and structure.

A.1 Base Prompts

- **Neutral (strength 1)** `Answer the question truthfully and concisely.`
  `Q: {question}`
  `A:`

A.2 Uncertainty-Aware Prompts

- **Allow Uncertainty (strength 2)** `If you are not sure, say 'I don't know'.`
  `Q: {question}`
  `A:`

- **Confidence-Gated (strength 3)** `Only answer if you are confident. Otherwise say 'I don't know'.`
  `Q: {question}`
  `A:`

A.3 Format-Constrained Prompts

- **One Sentence (strength 2)** `Answer in exactly one sentence.`
  `Q: {question}`
  `A:`

- **JSON Output (strength 4)** `Return JSON with keys answer and confidence (0-1).`
  `Question: {question}`
  `JSON:`

A.4 Reasoning Prompts

- **Brief Reasoning (strength 3)** `Think briefly, then give the final answer.`
  `Q: {question}`
  `A:`

In total, we use six prompt templates spanning four families (base, uncertainty-aware, format-constrained, and reasoning), with varying levels of constraint and instruction strength.

B Free-form Evaluation Details

We evaluate free-form generations using an LLM-based judge. For each model response, the judge receives the question, a reference answer, and the generated answer, and returns a structured label indicating whether the response is `correct`, `incorrect`, or `unknown`.

B.1 Judge Prompt

The system prompt used for evaluation is shown below:

```
You are a strict factuality judge for
question answering.

Decide whether the model answer is
factually correct with respect to the
question and the reference answer.

Return ONLY valid JSON with keys: -
label: one of ["correct", "incorrect",
"unknown"] - confidence: number between
0 and 1 - reason: short explanation

Rules: - "correct" if the answer is
factually consistent with the reference,
including paraphrases. - "incorrect" if
it is wrong or contradicts the reference.
- "unknown" if it is too ambiguous,
evasive, malformed, or incomplete to
judge confidently. - Be tolerant of short
factual answers like names and dates.
```

B.2 Structured Output Format

The judge is constrained to return a structured JSON object with three required fields:

- `label`: one of `correct`, `incorrect`, or `unknown`
- `confidence`: a scalar between 0 and 1
- `reason`: a brief natural-language justification

We use deterministic decoding (temperature = 0) to ensure consistent judgments across repeated evaluations.

B.3 Label Mapping

For analysis, we map judge outputs to binary hallucination indicators by treating `incorrect` responses as hallucinations. Cases labeled `unknown` are excluded from some analyses but are rare in practice.

C Additional Results

C.1 Confidence Intervals

We report bootstrap 95% confidence intervals for flip rates across all models and datasets. Intervals are consistently narrow, supporting the statistical robustness of the observed instability patterns.

HaluEval

- Qwen-7B: 0.192 [0.184, 0.199]
- Phi-3-mini: 0.217 [0.209, 0.225]
- LLaMA-3.1-8B: 0.225 [0.217, 0.233]
- Mistral-7B: 0.369 [0.360, 0.379]
- Zephyr-7B: 0.598 [0.588, 0.606]

TruthfulQA

- Qwen-7B: 0.324 [0.293, 0.357]
- Phi-3-mini: 0.324 [0.293, 0.357]
- LLaMA-3.1-8B: 0.424 [0.390, 0.458]
- Mistral-7B: 0.452 [0.414, 0.486]
- Zephyr-7B: 0.580 [0.546, 0.613]