When Format Changes Meaning: Investigating Semantic Inconsistency of Large Language Models

Cheongwoong Kang[1], **Jongeun Baek**[1], **Yeonjea Kim**[1], **Jaesik Choi**[1,2]
[1]KAIST, [2]INEEJI
{cw.kang, jongeun.baek, yeon.kim, jaesik.choi}@kaist.ac.kr

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks. However, they remain vulnerable to semantic inconsistency, where minor formatting variations result in divergent predictions for semantically equivalent inputs. Our comprehensive evaluation reveals that this brittleness persists even in state-of-the-art models such as GPT-4o, posing a serious challenge to their reliability. Through a mechanistic analysis, we find that semantic-equivalent input changes induce instability in internal representations, ultimately leading to divergent predictions. This reflects a deeper structural issue, where form and meaning are intertwined in the embedding space. We further demonstrate that existing mitigation strategies, including direct fine-tuning on format variations, do not fully address semantic inconsistency, underscoring the difficulty of the problem. Our findings highlight the need for deeper mechanistic understanding to develop targeted methods that improve robustness.

1 Introduction

Large language models (LLMs) have become the foundation of modern natural language processing, achieving state-of-the-art performance across a wide range of tasks (Abdin et al., 2024; Dubey et al., 2024; Achiam et al., 2023). Despite their impressive capabilities, LLMs frequently exhibit semantic inconsistency, where minor variations lead to inconsistent predictions for semantically equivalent inputs (Sclar et al., 2024; Qi et al., 2023; Jang et al., 2022; Elazar et al., 2021; Zhao et al., 2021; Jin et al., 2020; Ravichander et al., 2020). As illustrated in Figure 1, even a frontier model like GPT-4o may produce different answers to the same question solely due to formatting changes. Such

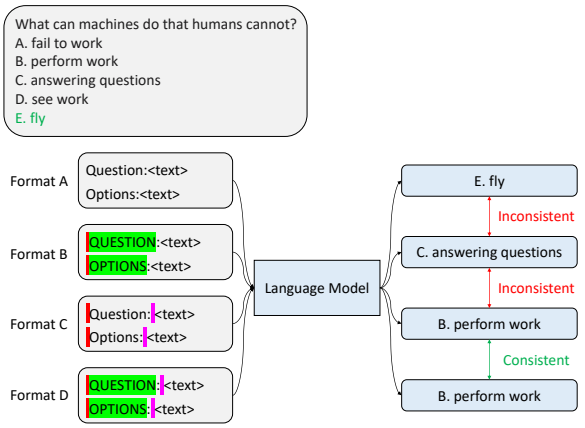


Figure 1: An example of inconsistent predictions of GPT-4o for the same question with slight format variations. The green blocks indicate a change in casing, while the red/pink vertical lines represent whitespaces.

inconsistencies raise concerns about the trustworthiness and reasoning stability of LLMs, suggesting that these models may not possess genuine semantic understanding but rather function as stochastic parrots, generating responses based on surface-level patterns (Kang and Choi, 2023; Joshi et al., 2022; Bender et al., 2021; Simon, 1954).

In this work, we systematically examine semantic inconsistency of LLMs, investigating whether they provide consistent predictions across semantically equivalent inputs. Our comprehensive evaluation across diverse tasks, model and prompting strategies quantifies the pervasiveness of issue, revealing that even the state-of-the-art models such as GPT-4o remain vulnerable. This persistent brittleness presents a serious obstacle for deploying LLMs in high-stakes domains such as law, finance and healthcare, where consistency and trust are essential.

To this end, our work aims to understand when and why semantic inconsistency emerges and how it might be controlled. We first provide a mechanistic diagnosis, tracing semantic inconsistency from

*Code and data are available at https://github.com/CheongWoong/RoLM.

superficial input perturbations to instability in internal embeddings, revealing a failure to form format-invariant representations. This analysis suggests that form and meaning are not clearly separated in the embedding space. We then show that common mitigation strategies, such as fine-tuning on diverse format variations, are insufficient to eliminate inconsistency, indicating that the problem extends beyond the input-output level. Finally, by connecting these failures to their internal representational causes, we motivate the need for representation-aware solutions that address the root mechanisms of inconsistency.

2 Related Work

2.1 Prompt Format Variations

Recent studies have shown that large language models (LLMs) are sensitive to superficial prompt format variations, changes in input structure that preserve the exact wording but differ in formatting, often resulting in significant performance variance (Sclar et al., 2024; Zhao et al., 2021). However, these works primarily highlight performance fluctuations, without directly assessing whether the model's predictions remain semantically consistent across formats. In contrast, our work focuses explicitly on semantic consistency under prompt format variations.

2.2 Semantic Consistency

Semantic consistency refers to the ability of a model to produce invariant predictions across semantically equivalent inputs (Jang et al., 2022). Prior work have evaluated consistency under various input perturbations such as paraphrasing (Jang et al., 2022; Elazar et al., 2021; Jin et al., 2020), syntactic rewrites (Ravichander et al., 2020) and translation (Wang et al., 2025; Qi et al., 2023), typically using pairwise metrics to assess consistency. Our work extends this line of research in two important directions. First, we study a previously underexplored class of perturbations, prompt format variations, which are particularly relevant in real-world deployment. Second, we introduce a setwise consistency metric that imposes a stricter criterion: models must remain consistent across all format variants, not just between pairs. This allows for a more principled and challenging evaluation of consistency.

2.3 Mechanistic Interpretability of Semantic Inconsistency

Beyond behavioral evaluation, recent work has begun exploring the internal mechanisms underlying semantic inconsistency. For example, Yang et al. (2025) and Yang et al. (2024) identify model components that contribute to inconsistent behavior and use activation steering techniques to mitigate it. While these studies focus on intervening to reduce inconsistency, our work contributes a novel diagnostic perspective. To the best of our knowledge, we are the first to validate a mechanistic chain that links input-level format distance to intermediate representational instability and finally to output-level inconsistency. This deeper mechanistic understanding goes beyond simple behavioral observations, identifying not only where inconsistency arises but also how it propagates through the model's internal computation. These insights, in turn, motivate the development of representation-level solutions that target the root causes of inconsistency.

3 Measuring Semantic Consistency under Prompt Format Variations

Understanding how language models respond to semantically equivalent inputs with different surface forms is crucial for diagnosing their generalization and abstraction capabilities. In this work, we focus on semantic consistency, which captures whether a model produces stable predictions across semantic-preserving input variations (Jang et al., 2022; Elazar et al., 2021). We specifically analyze prompt format variations (Sclar et al., 2024; Zhao et al., 2021), perturbations that alter input formatting (e.g., spacing, capitalization) while preserving the exact wording of the main content verbatim. This choice is motivated by the prevalence of such variations in real-world use, where users may unintentionally modify prompt formats without altering the semantic content.

We propose a novel evaluation framework that combines: (1) setwise consistency, a stricter metric that requires identical outputs across all semantic-preserving variants of an input, and (2) prompt format variations, a class of overlooked yet practically important perturbations.

3.1 Semantic Consistency Metrics

To assess semantic consistency, we use both pairwise and setwise metrics. Pairwise consistency

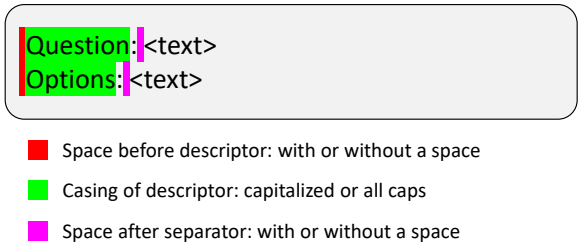


Figure 2: An example of prompt format variations.

measures whether a model returns the same prediction for two semantically equivalent inputs X_j and X_k (Elazar et al., 2021):

$$\text{Consistency}_{\text{pairwise}}(j,k) = \mathbb{1}(\hat{y}_j = \hat{y}_k) \quad (1)$$

where $\mathbb{1}$ is the indicator function.

While pairwise consistency offers a localized view, it cannot detect fragmented inconsistencies across a set of inputs. To address this, we introduce setwise consistency, which evaluates whether a model produces the same output across an entire set $\mathcal{S}$ of semantically equivalent variations:

$$\text{Consistency}_{\text{setwise}}(\mathcal{S}) = \mathbb{1}\left(|\{\hat{y}_k \mid k \in \mathcal{S}\}| = 1\right) \quad (2)$$

This returns 1 if and only if all predictions in the set are identical. Notably, pairwise consistency is a special case of setwise consistency when $|\mathcal{S}| = 2$.

We define semantic inconsistency as $1 -$ consistency. Higher inconsistency indicates greater model sensitivity to semantically equivalent perturbations, suggesting instability in the model behavior. Throughout the paper (Sections 4.2, 5.3 and 6), we use setwise consistency as our primary evaluation criterion, while pairwise consistency supports for more fine-grained analyses.

3.2 Format Variation Design

Inspired by Sclar et al. (2024), we examine three common and impactful prompt format dimensions: (1) spacing before descriptors (present vs. absent), (2) casing of descriptors (capitalized vs. all caps) and (3) separator spacing (present vs. absent). Each binary option yields $2^3 = 8$ total format variants per input. While our experiments focus on this curated set, our framework is broadly applicable to other types of interventions.

4 Experiments

4.1 Setup

4.1.1 Datasets

We evaluate semantic consistency using five diverse datasets spanning commonsense, multi-hop reasoning, fact verification, mathematical problem solving and legal reasoning. CommonsenseQA (Talmor et al., 2019) is a multiple-choice question answering (MCQA) dataset that requires commonsense knowledge, where each question has five answer options. QASC (Khot et al., 2020) is an MCQA dataset focused on multi-hop reasoning over scientific facts. 100TFQA [1] is a collection of true/false factual questions curated from a public domain website. GSM8K (Cobbe et al., 2021) contains various grade-school math word problems that require step-by-step numerical reasoning. MMLU-Pro-Law-100Q includes 100 multiple-choice questions from the Law domain of the MMLU-Pro (Wang et al., 2024).

To ensure robust evaluation, we split each dataset into validation and test sets with a 20:80 ratio, using the validation set to sample few-shot demonstrations. Detailed dataset statistics and input examples are provided in Appendix A.

4.1.2 Input Construction

For each input sample, we construct eight formatting variants using all binary combinations of three formatting dimensions. We then compute pairwise and setwise consistency scores by comparing model outputs across these variants.

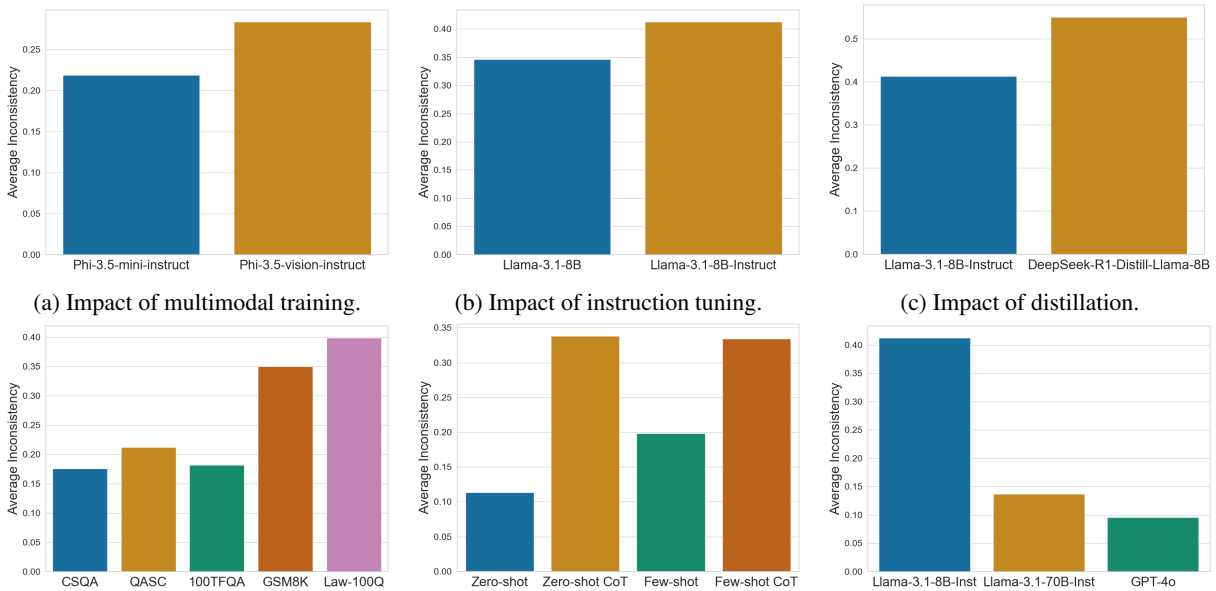
4.1.3 Models

We evaluate a range of language models, including Phi-3.5-mini-instruct, Phi-3.5-vision-instruct (Abdin et al., 2024), Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Dubey et al., 2024), DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) and GPT-4o (Achiam et al., 2023). These selections allow for targeted comparisons across architecture, instruction tuning, distillation and model scale. To ensure reproducibility, we use greedy decoding across all models.

4.1.4 Prompting Strategies

We access four prompting strategies: (1) zero-shot, (2) zero-shot chain-of-thought (CoT) (Kojima et al., 2022), (3) few-shot (Brown et al., 2020) and (4) few-shot CoT (Wei et al., 2022). Each method probes a different axis of model behavior. Zero-shot prompts test foundational capabilities, while CoT encourages intermediate reasoning. Few-shot variants provide in-context demonstrations to guide behavior. In the few-shot settings, we provide four demonstrations per input.

[1] 100 True or False Questions



(a) Impact of multimodal training. (b) Impact of instruction tuning. (c) Impact of distillation.

(d) Comparison between tasks. (e) Comparison between prompting. (f) Impact of model scaling.

Figure 3: Impact of various modeling, prompting and task-related factors on semantic inconsistency.

4.2 Results

Figure 3 summarizes average inconsistency across different settings. A complete breakdown of results by task and model is provided in Appendix C. Figure 3 illustrates the impact of various model design choices on semantic inconsistency. We find that multimodal training, instruction tuning and distillation (Figures 3a–3c) do not reduce inconsistency. Across tasks (Figure 3d), inconsistency is most severe in domains requiring complex reasoning, such as professional legal questions (MMLU-Pro-Law-100Q) and mathematical word problems (GSM8K). This suggests that complex or domain-specific reasoning amplifies the model's sensitivity to surface-level variations. The highest inconsistency rate was observed in MMLU-Pro-Law-100Q, confirming that this issue becomes more pronounced in more complex and realistic settings.

Prompting strategies (Figure 3e), including few-shot and CoT methods, do not improve consistency. In contrast, model scale emerges as a major factor (Figure 3f). For instance, Llama-3.1-70B-Instruct consistently outperforms its 8B counterpart. While GPT-4o achieves the highest overall consistency, Llama-3.1-70B-Instruct surpasses it on certain datasets (Table 9), indicating that factors beyond size, such as architecture and training regimen, also play an important role.

In summary, model scale correlates with improved consistency, but even frontier models like GPT-4o exhibit persistent inconsistency under minor prompt changes. This persistent failure raises open questions about the current limits of large language model (LLM) architectures and their ability to generalize meaning beyond surface-level cues (Bender and Koller, 2020).

4.3 Human Evaluation

To establish a reference point for model behavior, we conducted a consistency experiment using the 100TFQA dataset with three graduate-level participants. The evaluation included two conditions: (1) Format Variation, where each question was shown in all eight prompt formats used in our main study, and (2) Repeat Prompts, a control condition where the same format was repeated eight times to isolate baseline inconsistency due to human factors such as fatigue or distraction.

The results show that humans are highly robust to superficial formatting changes. Two participants (P1 and P2) demonstrated near-perfect consistency, with inconsistency rates of 0.000 and 0.013 under the Format Variation setting. The third participant (P3) exhibited higher inconsistency (0.163), but also showed a high inconsistency rate (0.238) in the control condition, suggesting that their inconsistency stemmed from transient attention lapses rather than format sensitivity.

These results highlight a key distinction: while human inconsistency may arise from cognitive factors, LLM inconsistency appears to be a systemic vulnerability in the model's internal representa-

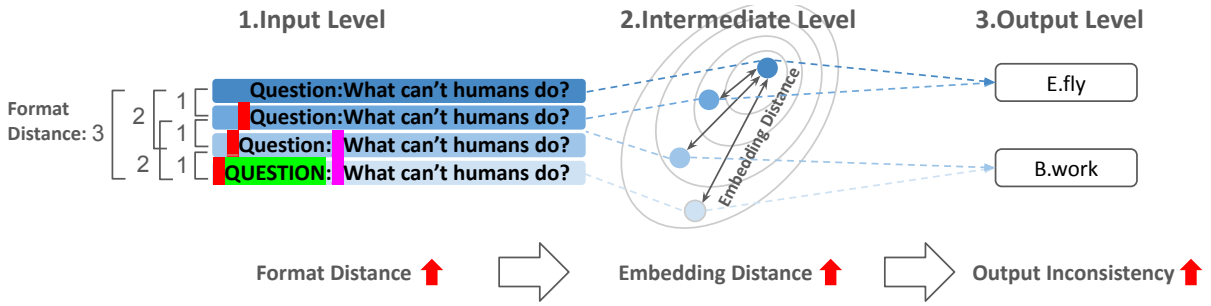


Figure 4: Conceptual illustration of the propagation of format-induced inconsistency. Superficial input-level variations (Format Edit Distance) lead to representational instability (Embedding Distance), which subsequently results in output inconsistency. This mechanism suggests that form and meaning are intertwined in the embedding space of language models.

tions. In this light, understanding the underlying mechanisms of semantic consistency is essential for building stable and trustworthy LLMs, especially in high-stakes applications.

5 Tracing the Propagation of Semantic Inconsistency

We present a mechanistic analysis tracing how superficial format variations propagate through a model's internal representations, ultimately leading to output inconsistencies. First, we establish the primary propagation chain: from input-level format differences to instabilities in intermediate representations and finally to output-level inconsistency, as summarized in Figure 4 (Section 5.1). We then examine the context-dependent impact of specific format elements (Section 5.2) and analyze how model confidence correlates with consistency (Section 5.3).

5.1 Format Variations Propagate Through Embedding to Output

To understand how prompt formats are encoded by the model, we begin by visualizing output embeddings from the final layer at the last token position of the input prompt. Figure 5 shows a 2D projection via PCA (Pearson, 1901) of these embeddings for Llama-3.1-8B-Instruct on CommonsenseQA under the zero-shot setting. Distinct formatting variants form separable clusters in the embedding space, suggesting that format differences are encoded in the model's internal representations. This observation is quantitatively supported by a linear probe trained on the top four principal components, which achieves over 99% accuracy in classifying the format variants.

To assess whether these embedding differences

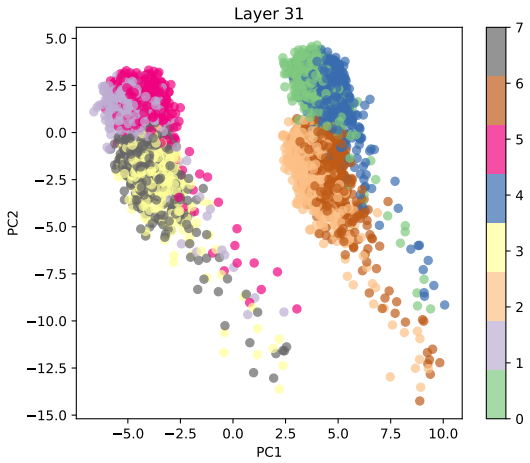


Figure 5: Distribution of last-layer embeddings for Llama-3.1-8B-Instruct on CommonsenseQA (zero-shot setting). Inputs are clustered according to their formatting, indicating that surface-level differences are encoded in the embedding space.

contribute to semantic inconsistency, we compute the Spearman's rank correlation (Spearman, 1904) between embedding distance (Euclidean) and pairwise consistency. For Llama-3.1-8B-Instruct, correlation coefficients in the most predictive layer range from -0.31 to -0.64 across tasks, indicating that larger embedding differences between format variants are associated with higher inconsistency in predictions. To further localize where this divergence occurs, we compute layer-wise correlations between embedding distance and pairwise consistency. As shown in Figure 6, the strongest correlations are observed in the middle layers (typically between layers 10 and 18), highlighting them as the primary locus of representational instability. Full results are provided in Appendix D.

Next, we examine how input-level format differ-

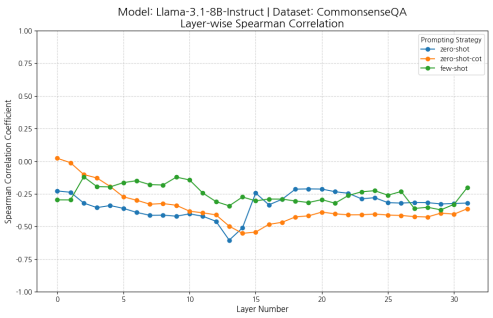


Figure 6: Layer-wise Spearman correlation between embedding distance and pairwise consistency of Llama-3.1-8B-Instruct on CommonsenseQA. The correlation peaks in the middle layers (around layer 13), suggesting that these layers are most sensitive to format-induced representational divergence.

ences propagate to embedding space. We introduce format edit distance, which measures the difference between two format variants using Hamming distance over formatting elements:

$$d(f_1, f_2) = \sum_{i=1}^{N} \mathbb{1}(f_{1,i}, f_{2,i}) \quad (3)$$

where N is the number of formatting dimensions. For Llama-3.1-8B-Instruct, the correlation between format edit distance and embedding distance ranges from 0.64 to 0.92 in the most predictive layer, again peaking in the middle layers (Appendix E).

Finally, to determine whether format edit distance directly impacts semantic inconsistency, we compute its correlation with pairwise consistency. Table 12 reports values between -0.40 and -0.60 for 12 out of 17 cases, showing a moderate negative correlation. These values closely mirror the correlations between embedding distance and pairwise consistency (–0.31 to –0.64), suggesting a coherent propagation chain. This propagation is summarized in Figure 4.

Our analysis provides a key diagnostic insight: semantic inconsistency is not merely a surface-level phenomenon, but reflects a deep representational failure within the model, where form and meaning are not properly disentangled. Our findings not only pinpoint where the problem arises, but also inform how future interpretability and robustness efforts should be targeted.

5.2 Context-Dependent Impact of Formatting Elements

To better understand the input-level drivers of semantic inconsistency, we isolate the effect of each formatting element and assess whether certain elements are inherently more disruptive than others. We begin by toggling each formatting element independently and measuring the average inconsistency it induces. Across tasks, models and prompting strategies, the three elements produce similar average inconsistency scores: 0.11, 0.11 and 0.10, respectively. This suggests that no single formatting element dominates in its effect on inconsistency. At a global level, the model appears similarly sensitive to all types of superficial changes.

Despite their similar overall impact, we observe that the relative effect of each element is not consistent across different settings. In some settings, a particular element induces the highest inconsistency, while in others it induces the least. To quantify this variability, we compute rank correlations between the inconsistency rankings of the three elements across controlled conditions: (A) varying tasks while fixing model and prompting strategy, (B) varying models while fixing task and prompting strategy and (C) varying prompting strategies while fixing model and task. The resulting average pairwise correlations are low: 0.18 (A), 0.17 (B) and 0.14 (C). These results indicate that the relative importance of formatting elements is highly context-sensitive. In other words, the same formatting change can have very different impacts depending on the task, model or prompting method in use. This context-dependence poses a major challenge for robustness. While one might hope to mitigate inconsistency by avoiding a particular problematic formatting element, our findings show that no such universal culprit exists. Instead, robust mitigation methods must account for the broader interaction between input features and model context, rather than relying on global heuristics.

5.3 Model Uncertainty as a Correlate of Inconsistency

Beyond input-level and representation-level factors, another potential factor explaining inconsistency lies at the output level: the model's uncertainty. To quantify this, we use the maximum softmax probability of the predicted answer as a proxy for confidence (Hendrycks and Gimpel, 2017). To examine its relationship with consistency, we compute Spearman correlation coefficients between confidence scores and setwise consistency. As shown in Table 13, Llama-3.1-8B-Instruct exhibits moderate positive correlations between 0.30 and 0.67 in 13 out of 17 settings. This finding is practically

Task	Model	Zero-shot	Zero-shot CoT	Few-shot	Few-shot CoT
CommonsenseQA	Llama-3.1-8B-Instruct	0.75 | 0.10	0.76 | 0.25	0.68 | 0.51	-
	+ Self-consistency	0.75 | 0.14	0.77 | 0.20	0.69 | 0.53	-
	+ Fine-tuning	0.79 | 0.09	0.77 | 0.25	0.77 | 0.13	-
	gpt-4o-2024-11-20	0.85 | 0.07	0.84 | 0.08	0.87 | 0.04	-
	+ Self-consistency	0.85 | 0.08	0.84 | 0.08	0.87 | 0.06	-
QASC	Llama-3.1-8B-Instruct	0.82 | 0.09	0.82 | 0.19	0.61 | 0.84	0.68 | 0.74
	+ Reference-guided	0.91 | 0.07	-	-	-
	+ Self-consistency	0.82 | 0.13	0.84 | 0.16	0.61 | 0.85	0.73 | 0.59
	+ Fine-tuning	0.84 | 0.10	0.84 | 0.19	0.76 | 0.33	0.73 | 0.59
	gpt-4o-2024-11-20	0.92 | 0.05	0.91 | 0.06	0.94 | 0.04	0.93 | 0.04
	+ Reference-guided	0.96 | 0.03	-	-	-
	+ Self-consistency	0.91 | 0.06	0.90 | 0.06	0.94 | 0.05	0.93 | 0.03
100TFQA	Llama-3.1-8B-Instruct	0.70 | 0.10	0.72 | 0.26	0.70 | 0.24	0.68 | 0.48
	+ Self-consistency	0.69 | 0.10	0.74 | 0.20	0.70 | 0.35	0.67 | 0.32
	+ Fine-tuning	0.69 | 0.10	0.71 | 0.28	0.67 | 0.34	0.65 | 0.42
	gpt-4o-2024-11-20	0.93 | 0.06	0.97 | 0.05	0.94 | 0.00	0.98 | 0.05
	+ Self-consistency	0.94 | 0.08	0.97 | 0.02	0.94 | 0.00	0.98 | 0.01
GSM8K	Llama-3.1-8B-Instruct	-	0.54 | 0.75	-	0.79 | 0.37
	+ Self-consistency	-	0.74 | 0.52	-	0.87 | 0.23
	+ Fine-tuning	-	0.66 | 0.52	-	0.65 | 0.48
	gpt-4o-2024-11-20	-	0.91 | 0.13	-	0.95 | 0.05
	+ Self-consistency	-	0.94 | 0.08	-	0.96 | 0.03

Table 1: Mean accuracy (left) and setwise inconsistency (right) are presented. Blue values indicate performance improvement over baseline (no mitigation), while red values denote performance drop.

significant, as it suggests that model confidence may serve as a lightweight, real-time indicator of potential inconsistency.

6 The Limits of Existing Mitigation Strategies

Our earlier analysis (Section 5) diagnosed semantic inconsistency as a deep representational failure. This motivates an evaluation of existing mitigation strategies to assess their effectiveness and limitations. We examine (1) prompt-based techniques that do not modify model parameters and (2) a more direct approach using fine-tuning with diverse format variations.

6.1 Prompt-Based Approaches

We first evaluate two prompt-based methods that operate without altering the model's internal weights. Reference-guided prompting incorporates key supporting information directly into the input to better anchor model reasoning. Unlike retrieval-augmented generation (Shi et al., 2024; Lewis et al., 2020), which retrieves external evidence, our setup uses ground-truth evidence to guide responses. We provide partial evidence (as shown in Table 5) from the ground-truth context to avoid reducing the task to simple answer extraction. Applied to QASC, where each question is supported by two facts, we concatenate a single supporting fact to the input. As shown in Table 1, reference-guided prompting improves both accuracy and consistency, but its practical utility is constrained by the availability of such evidence in real-world settings.

Self-consistency is an output-based ensemble technique that generates multiple responses for the same input and selects the most frequent answer via majority voting (Wang et al., 2023). This approach assumes that the most consistent response across multiple trials represents the most reliable answer. We produce 10 responses per input using nucleus sampling (Holtzman et al., 2020) with a top_p value of 0.9. While Table 1 shows meaningful gains, especially in chain-of-thought (CoT) settings, its effectiveness diminishes in non-CoT

settings, where inconsistency can even increase in some cases. Additionally, self-consistency is computationally expensive due to repeated sampling, which limits its scalability in real-world deployments.

6.2 Fine-Tuning with Diverse Formats

We next consider a more direct approach: fine-tuning the model with diverse formats. Specifically, we fine-tune Llama-3.1-8B-Instruct on validation sets from four tasks, expanded to include all eight formatting variations per sample. Note that the model is trained for each task, separately.

We evaluate the fine-tuned model under two test conditions: (1) on unseen inputs with seen format types and (2) on seen inputs with unseen format types. In the first setting, fine-tuning reduces inconsistency, as shown in Table 1. However, in the second setting, results from CommonsenseQA (zero-shot) with 32 unseen format types (used in Appendix H) show that inconsistency actually increases (from 0.18 to 0.21), despite a substantial boost in task accuracy (from 0.76 to 0.92). These results suggest that fine-tuning struggles to generalize to new formatting combinations.

6.3 Diagnosis and The Path Forward

These findings demonstrate that the existing methods fall short of fully addressing semantic inconsistency. Prompt-based techniques show limited gains and rely on conditions (e.g., evidence access, sampling overhead) that constrain practical use. Fine-tuning on diverse formats, meanwhile, fails to generalize to unseen combinations, supporting our earlier diagnosis that inconsistency is not merely a surface-level problem but a deeper representational issue. This reinforces the need for principled approaches that target internal representations. Future work should focus on objectives that align intermediate representations across semantic-preserving variations, paving the way for more robust and trustworthy systems.

7 Conclusion

This study systematically investigates semantic consistency of large language models (LLMs), showing that even state-of-the-art models like GPT-4o remain vulnerable to minor prompt format variations. Our mechanistic analysis moves beyond behavioral observation to diagnose the root cause, tracing the problem to a deep representational failure, where form and meaning are intertwined. We further demonstrate that widely used mitigation strategies, including prompt-based methods and fine-tuning, do not fully address the problem. Taken together, our results motivate deeper mechanistic understanding to address the root causes of semantic inconsistency.

Limitations

Our analysis is limited in scope due to the high computational demands of probing large-scale models across multiple configurations. Specifically, we restrict our evaluation to a curated set of prompt format variations, models and tasks. Expanding this framework to a broader set of perturbations and domains would help assess the generality of our findings. Additionally, our embedding-level analysis provides only a coarse-grained view of internal dynamics. Factors such as superposition and polysemanticity limit the interpretability of aggregated representation distances. While our results offer a high-level diagnostic of representational failure, a deeper mechanistic understanding remains an open and promising direction for future work. By acknowledging these limitations, we hope to set a foundation for more precise diagnostic tools and more effective mitigation strategies that can improve the semantic robustness of LLMs in real-world deployments.

Acknowledgements

This work was partly supported by Institute for Information & communications Technology Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Support Program (KAIST)), (RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), (RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics), (RS-2024-00457882, AI Research Hub Project), (RS-2024-00509258, AI Guardians: Development of Robust, Controllable, and Unbiased Trustworthy AI Technology).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach,

Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP (Findings of EMNLP)*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM)*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Herbert A Simon. 1954. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association (JASA)*.

C Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology (AJP)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2024. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. In *Findings of the Association for Computational Linguistics: ACL (Findings of ACL)*.

Jingyuan Yang, Rongjun Li, Weixuan Wang, Ziyu Zhou, Zhiyong Feng, and Wei Peng. 2025. Lf-steering: Latent feature activation steering for enhancing semantic consistency in large language models. *arXiv preprint arXiv:2501.11036*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

A Dataset Descriptions

A.1 Dataset Statistics

Dataset	Domain	Answer Format	Validation Samples	Test Samples
CommonsenseQA	Commonsense	Multiple Choice (5)	244	977
QASC	Science	Multiple Choice (8)	185	741
100TFQA	Factual	True/False	20	80
GSM8K	Math	Number	263	1056
MMLU-Pro-Law-100Q	Law	Multiple Choice (10)	5	100

Table 2: Dataset statistics.

A.2 Input Examples

This section shows input examples used in the experiments. Task instructions are shown in Table 3 and input examples are shown in Table 4. The final input is formatted as "{task_instruction}\n\n{task_input}."

Task	Setting	Example
Multiple-choice QA	w/o CoT	Answer the following multiple-choice questions. Select the best answer from the given options and provide your output in the following valid JSON format: "'json {"Answer":"{letter}"}"' Do not include any additional text.
	CoT	Answer the following multiple-choice questions. Think step-by-step and provide a concise reasoning process that justifies your answer. Based on the reasoning, select the best answer from the given options and provide your output in the following valid JSON format: "'json {"Explanation":"{concise reasoning}", "Answer":"{letter}"}"' Ensure the explanation is minimal sufficient. Do not include any additional text.
	Reference-guided	Answer the following multiple-choice questions. Based on the provided reference, select the best answer from the given options and provide your output in the following valid JSON format: "'json {"Answer": "{letter}"}"' Do not include any additional text.
True/False QA	w/o CoT	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: "'json {"Answer":"{True/False}"}"' Do not include any additional text.
	CoT	Answer the following true or false questions. Think step-by-step and provide a concise reasoning process that justifies your answer. Based on the reasoning, determine whether the statement is True or False and provide your output in the following valid JSON format: "'json {"Explanation":"{concise reasoning}", "Answer":"{True/False}"}"' Ensure the explanation is minimal sufficient. Do not include any additional text.
Mathmatics	CoT	Answer the following math questions. Think step-by-step and provide a concise reasoning process that justifies your answer. Based on the reasoning, compute the correct numerical answer and provide your output in the following valid JSON format: "'json {"Explanation":"{concise reasoning}", "Answer":"{numeric answer}"}"' Ensure the explanation is minimal sufficient. Ensure that the answer is a pure number without any symbols, units, or explanations. Do not include any additional text.

Table 3: Examples of task instructions.

Task	Example
Multiple-choice QA	Question:From where does a snowflake form? Options: A. cloud B. snow storm C. billow D. air E. snowstorm
True/False QA	Statement:Alaska has the most active volcanoes of any state in the United States.
Mathematics	Question:Finley went to the grocery store and bought rice, beans, and pork for use in their home. It took her 20 more minutes to cook pork than rice, while beans took half the combined cooking time of pork and rice. If it took her 30 minutes to cook rice, how long in minutes did it take to cook all the food?

Table 4: Zero-shot task input examples.

Example
Question:what is saturated fat at room temperature? Options: A. solid B. cats C. cows D. steak E. gas F. liquid G. Protein H. unsaturated Reference:Butter is a fat that is a solid at room temperature.

Table 5: Reference-guided input examples from QASC.

A.3 Input Examples with Prompt Format Variations

Format index	Space after separator	Descriptor casing	Space before descriptor
0	No	Capitalized	No
1	No	Capitalized	Yes
2	No	All caps	No
3	No	All caps	Yes
4	Yes	Capitalized	No
5	Yes	Capitalized	Yes
6	Yes	All caps	No
7	Yes	All caps	Yes

Table 6: Prompt format variations.

Format index	Example
0	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"Answer":"{True/False}"}“‘ Do not include any additional text. Statement:Alaska has the most active volcanoes of any state in the United States.
1	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"Answer":"{True/False}"}“‘ Do not include any additional text. Statement:Alaska has the most active volcanoes of any state in the United States.
2	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"ANSWER":"{True/False}"}“‘ Do not include any additional text. STATEMENT:Alaska has the most active volcanoes of any state in the United States.
3	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"ANSWER":"{True/False}"}“‘ Do not include any additional text. STATEMENT:Alaska has the most active volcanoes of any state in the United States.
4	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"Answer":"{True/False}"}“‘ Do not include any additional text. Statement:Alaska has the most active volcanoes of any state in the United States.
5	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"Answer":"{True/False}"}“‘ Do not include any additional text. Statement:Alaska has the most active volcanoes of any state in the United States.
6	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"ANSWER":"{True/False}"}“‘ Do not include any additional text. STATEMENT:Alaska has the most active volcanoes of any state in the United States.
7	Answer the following true or false questions. Determine whether the statement is True or False and provide your output in the following valid JSON format: “‘json {"ANSWER":"{True/False}"}“‘ Do not include any additional text. STATEMENT:Alaska has the most active volcanoes of any state in the United States.

Table 7: Format variation examples from 100TFQA (zero-shot).

B Model Specifications

Model	Parameters	Instruction tuning	Multimodal training
Phi-3.5-mini-instruct	3.8B	Yes	No
Phi-3.5-vision-instruct	4.2B	Yes	Yes
Llama-3.1-8B	8B	No	No
Llama-3.1-8B-Instruct	8B	Yes	No
Llama-3.1-70B-Instruct	70B	Yes	No
DeepSeek-R1-Distill-Llama-8B	8B	Yes	No
GPT-4o*	-	Yes	Yes

* gpt-4o-2024-11-20

Table 8: Model details.

C Detailed Results

It is important to note certain data availability limitations: Few-shot CoT results for CommonsenseQA are not presented due to the unavailability of labeled explanations. For GSM8K, non-CoT prompting results were excluded as models demonstrated poor performance without CoT prompting. Additionally, Llama-3.1-8B's zero-shot results are unavailable due to its inability to consistently follow task instructions in a purely zero-shot setting. Similarly, DeepSeek-R1-Distill-Llama-8B inherently generates step-by-step CoT reasoning even when not explicitly prompted for it, leading to the omission of its non-CoT results.

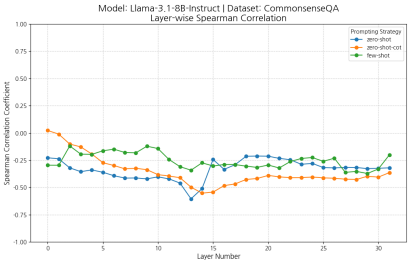
Task	Model	Zero-shot	Zero-shot CoT	Few-shot	Few-shot CoT
CommonsenseQA	Phi-3.5-mini-instruct	0.75 | 0.10	0.76 | 0.21	0.73 | 0.10	-
	Phi-3.5-vision-instruct	0.76 | 0.09	0.73 | 0.30	0.74 | 0.10	-
	Llama-3.1-8B	-	-	0.72 | 0.11	-
	Llama-3.1-8B-Instruct	0.75 | 0.10	0.76 | 0.25	0.68 | 0.51	-
	Llama-3.1-70B-Instruct	0.82 | 0.04	0.84 | 0.12	0.80 | 0.26	-
	DeepSeek-R1-Distill-Llama-8B	-	0.68 | 0.51	-	-
	gpt-4o-2024-11-20	0.85 | 0.07	0.84 | 0.08	0.87 | 0.04	-
QASC	Phi-3.5-mini-instruct	0.76 | 0.09	0.79 | 0.16	0.77 | 0.10	0.82 | 0.17
	Phi-3.5-vision-instruct	0.75 | 0.11	0.77 | 0.29	0.76 | 0.10	0.77 | 0.20
	Llama-3.1-8B	-	-	0.78 | 0.10	0.69 | 0.32
	Llama-3.1-8B-Instruct	0.82 | 0.09	0.82 | 0.19	0.61 | 0.84	0.68 | 0.74
	Llama-3.1-70B-Instruct	0.91 | 0.04	0.92 | 0.05	0.90 | 0.21	0.92 | 0.05
	DeepSeek-R1-Distill-Llama-8B	-	0.69 | 0.52	-	0.71 | 0.52
	gpt-4o-2024-11-20	0.92 | 0.05	0.91 | 0.06	0.94 | 0.04	0.93 | 0.04
100TFQA	Phi-3.5-mini-instruct	0.66 | 0.11	0.68 | 0.14	0.69 | 0.05	0.65 | 0.14
	Phi-3.5-vision-instruct	0.61 | 0.12	0.67 | 0.34	0.63 | 0.20	0.66 | 0.18
	Llama-3.1-8B	-	-	0.73 | 0.10	0.67 | 0.32
	Llama-3.1-8B-Instruct	0.70 | 0.10	0.72 | 0.26	0.70 | 0.24	0.68 | 0.48
	Llama-3.1-70B-Instruct	0.88 | 0.02	0.82 | 0.14	0.87 | 0.14	0.82 | 0.09
	DeepSeek-R1-Distill-Llama-8B	-	0.66 | 0.48	-	0.68 | 0.55
	gpt-4o-2024-11-20	0.93 | 0.06	0.97 | 0.05	0.94 | 0.00	0.98 | 0.05
GSM8K	Phi-3.5-mini-instruct	-	0.52 | 0.68	-	0.84 | 0.19
	Phi-3.5-vision-instruct	-	0.54 | 0.71	-	0.73 | 0.32
	Llama-3.1-8B	-	-	-	0.50 | 0.48
	Llama-3.1-8B-Instruct	-	0.54 | 0.75	-	0.79 | 0.37
	Llama-3.1-70B-Instruct	-	0.88 | 0.21	-	0.94 | 0.06
	DeepSeek-R1-Distill-Llama-8B	-	0.78 | 0.33	-	0.86 | 0.28
	gpt-4o-2024-11-20	-	0.91 | 0.13	-	0.95 | 0.05
MMLU-Pro-Law-100Q	Phi-3.5-mini-instruct	0.27 | 0.25	0.32 | 0.49	0.24 | 0.23	0.31 | 0.52
	Phi-3.5-vision-instruct	0.23 | 0.30	0.24 | 0.58	0.25 | 0.19	0.24 | 0.71
	Llama-3.1-8B	-	-	0.24 | 0.24	0.26 | 0.39
	Llama-3.1-8B-Instruct	0.29 | 0.29	0.30 | 0.65	0.17 | 0.64	0.20 | 0.53
	Llama-3.1-70B-Instruct	0.41 | 0.05	0.45 | 0.36	0.40 | 0.06	0.40 | 0.43
	DeepSeek-R1-Distill-Llama-8B	-	0.18 | 0.87	-	0.21 | 0.90
	gpt-4o-2024-11-20	0.58 | 0.18	0.57 | 0.26	0.58 | 0.17	0.56 | 0.28

Table 9: A comprehensive overview of average accuracy (left) and setwise inconsistency (right).

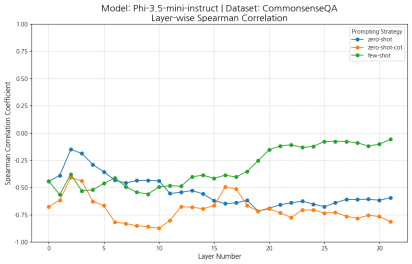
D Correlation between Embedding Distance and Output Consistency

Model	Dataset	Strategy	Mean	Std	Min (Best)	Max (Worst)	Best Layer
Phi-3.5-mini-instruct	CommonsenseQA	Zero-shot	-0.53	0.14	-0.72	-0.15	19
		Zero-shot CoT	-0.71	0.11	-0.88	-0.41	10
		Few-shot	-0.32	0.18	-0.57	-0.06	1
	QASC	Zero-shot	-0.36	0.13	-0.52	0.03	19
		Zero-shot CoT	-0.58	0.11	-0.69	-0.28	21
		Few-shot	-0.59	0.09	-0.74	-0.44	17
		Few-shot CoT	-0.24	0.04	-0.31	-0.16	31
	100TFQA	Zero-shot	-0.41	0.16	-0.87	-0.17	1
		Zero-shot CoT	-0.37	0.15	-0.57	-0.08	26
		Few-shot	-0.20	0.09	-0.39	-0.01	9
		Few-shot CoT	0.16	0.08	-0.05	0.28	0
	GSM8K	Zero-shot CoT	-0.36	0.14	-0.48	0.05	13
		Few-shot CoT	-0.50	0.07	-0.61	-0.35	18
	MMLU-Pro-Law-100Q	Zero-shot	-0.10	0.16	-0.29	0.25	20
		Zero-shot CoT	-0.41	0.08	-0.54	-0.22	21
		Few-shot	-0.14	0.12	-0.30	0.14	30
		Few-shot CoT	-0.63	0.18	-0.81	-0.06	16
Llama-3.1-8B-Instruct	CommonsenseQA	Zero-shot	-0.33	0.09	-0.60	-0.21	13
		Zero-shot CoT	-0.36	0.14	-0.55	0.02	14
		Few-shot	-0.26	0.07	-0.37	-0.12	29
	QASC	Zero-shot	-0.25	0.12	-0.56	-0.11	13
		Zero-shot CoT	-0.40	0.13	-0.61	-0.18	13
		Few-shot	-0.33	0.12	-0.52	-0.14	12
		Few-shot CoT	-0.20	0.14	-0.34	0.07	22
	100TFQA	Zero-shot	-0.11	0.12	-0.34	0.05	13
		Zero-shot CoT	-0.12	0.08	-0.33	-0.03	0
		Few-shot	-0.36	0.07	-0.49	-0.18	9
		Few-shot CoT	-0.31	0.05	-0.42	-0.14	2
	GSM8K	Zero-shot CoT	-0.11	0.10	-0.35	0.01	13
		Few-shot CoT	-0.41	0.10	-0.60	-0.23	29
	MMLU-Pro-Law-100Q	Zero-shot	-0.16	0.14	-0.46	0.03	10
		Zero-shot CoT	-0.46	0.11	-0.64	-0.29	11
		Few-shot	-0.19	0.07	-0.31	-0.03	18
		Few-shot CoT	-0.26	0.11	-0.38	-0.00	18

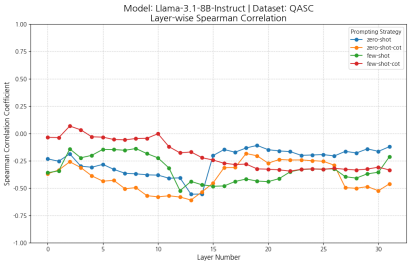
Table 10: Aggregation statistics of Spearman correlation coefficients between embedding distance and pairwise consistency for all layers.



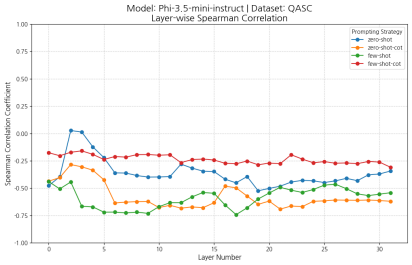
(a) Llama-3.1-8B-Instruct (CommonsenseQA)



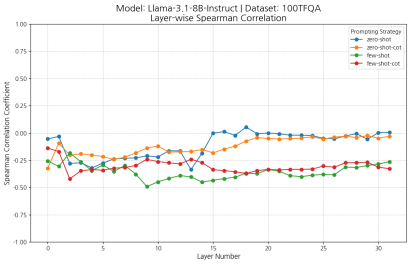
(b) Phi-3.5-mini-instruct (CommonsenseQA)



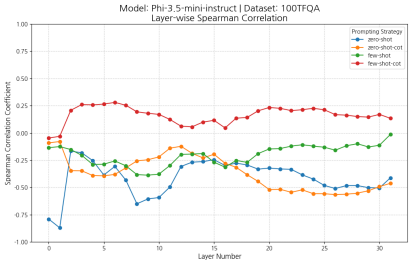
(c) Llama-3.1-8B-Instruct (QASC)



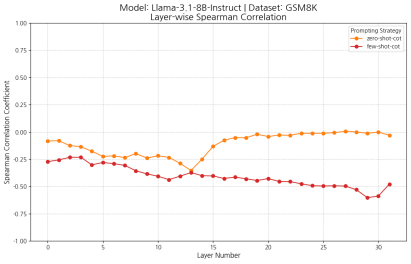
(d) Phi-3.5-mini-instruct (QASC)



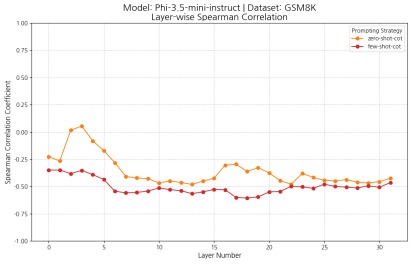
(e) Llama-3.1-8B-Instruct (100TFQA)



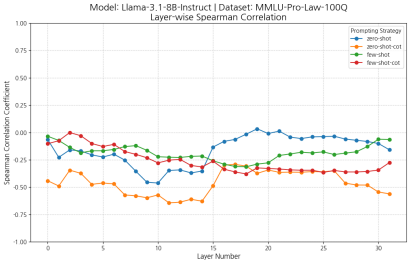
(f) Phi-3.5-mini-instruct (100TFQA)



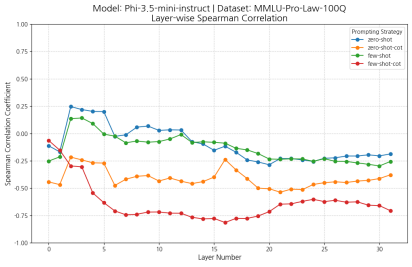
(g) Llama-3.1-8B-Instruct (GSM8K)



(h) Phi-3.5-mini-instruct (GSM8K)



(i) Llama-3.1-8B-Instruct (MMLU-Pro-Law-100Q)



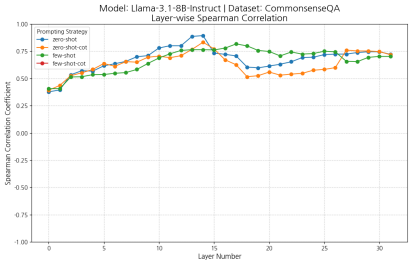
(j) Phi-3.5-mini-instruct (MMLU-Pro-Law-100Q)

Figure 7: Layer-wise Spearman correlation coefficients between embedding distance and pairwise consistency across tasks.

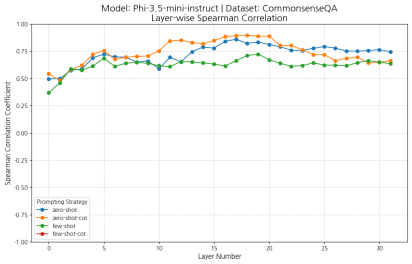
E Correlation between Format Edit Distance and Embedding Distance

Model	Dataset	Strategy	Mean	Std	Min (Worst)	Max (Best)	Best Layer
Phi-3.5-mini-instruct	CommonsenseQA	Zero-shot	0.72	0.09	0.50	0.86	17
		Zero-shot CoT	0.74	0.11	0.48	0.90	18
		Few-shot	0.62	0.06	0.37	0.72	19
	QASC	Zero-shot	0.71	0.10	0.44	0.84	17
		Zero-shot CoT	0.73	0.11	0.51	0.90	16
		Few-shot	0.64	0.08	0.40	0.78	17
		Few-shot CoT	0.63	0.07	0.45	0.76	19
	100TFQA	Zero-shot	0.82	0.11	0.50	0.90	19
		Zero-shot CoT	0.76	0.13	0.30	0.91	18
		Few-shot	0.73	0.09	0.47	0.87	11
		Few-shot CoT	0.67	0.06	0.51	0.78	13
	GSM8K	Zero-shot CoT	0.77	0.14	0.39	0.91	20
		Few-shot CoT	0.65	0.09	0.36	0.74	19
	MMLU-Pro-Law-100Q	Zero-shot	0.67	0.12	0.33	0.83	18
		Zero-shot CoT	0.69	0.13	0.40	0.92	17
		Few-shot	0.64	0.13	0.24	0.79	28
		Few-shot CoT	0.66	0.11	0.30	0.80	31
Llama-3.1-8B-Instruct	CommonsenseQA	Zero-shot	0.68	0.11	0.38	0.89	14
		Zero-shot CoT	0.63	0.10	0.39	0.83	14
		Few-shot	0.67	0.11	0.41	0.82	17
	QASC	Zero-shot	0.69	0.11	0.38	0.91	14
		Zero-shot CoT	0.60	0.11	0.43	0.78	14
		Few-shot	0.61	0.08	0.38	0.72	29
		Few-shot CoT	0.55	0.04	0.42	0.64	11
	100TFQA	Zero-shot	0.77	0.11	0.54	0.92	5
		Zero-shot CoT	0.68	0.14	0.52	0.92	3
		Few-shot	0.68	0.11	0.49	0.90	6
		Few-shot CoT	0.67	0.11	0.46	0.90	4
	GSM8K	Zero-shot CoT	0.57	0.12	0.44	0.88	13
		Few-shot CoT	0.58	0.06	0.41	0.71	31
	MMLU-Pro-Law-100Q	Zero-shot	0.70	0.12	0.41	0.90	14
		Zero-shot CoT	0.67	0.11	0.46	0.86	10
		Few-shot	0.65	0.08	0.42	0.75	18
		Few-shot CoT	0.60	0.07	0.43	0.77	10

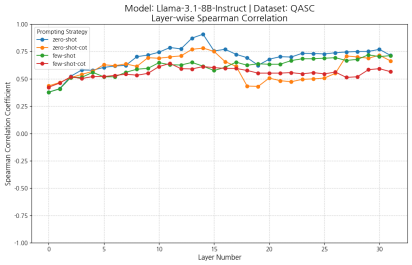
Table 11: Aggregation statistics of Spearman correlation coefficients between format edit distance and embedding distance for all layers.



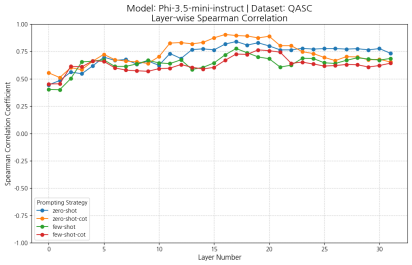
(a) Llama-3.1-8B-Instruct (CommonsenseQA)



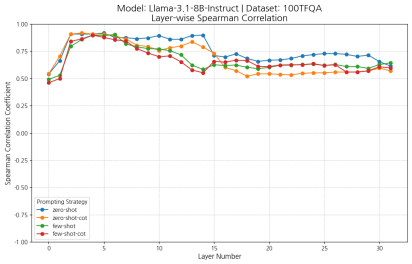
(b) Phi-3.5-mini-instruct (CommonsenseQA)



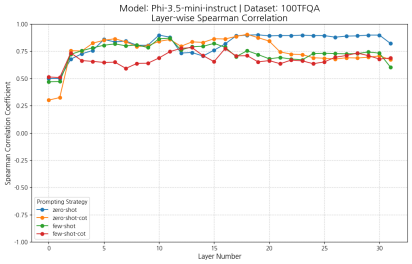
(c) Llama-3.1-8B-Instruct (QASC)



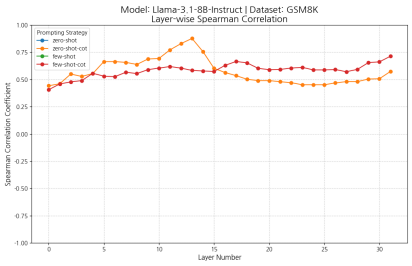
(d) Phi-3.5-mini-instruct (QASC)



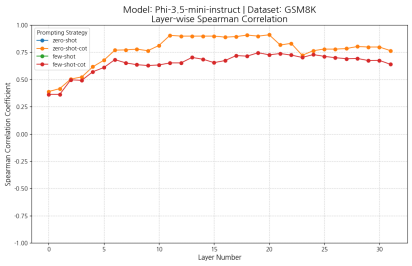
(e) Llama-3.1-8B-Instruct (100TFQA)



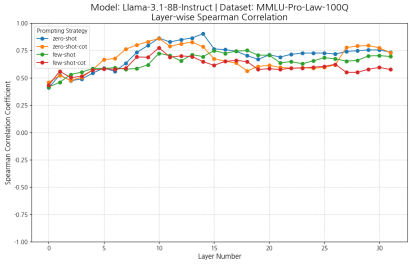
(f) Phi-3.5-mini-instruct (100TFQA)



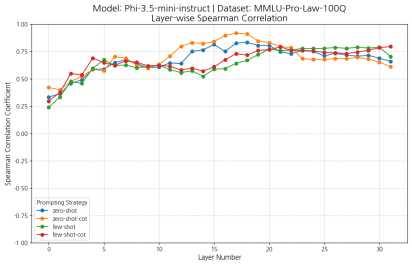
(g) Llama-3.1-8B-Instruct (GSM8K)



(h) Phi-3.5-mini-instruct (GSM8K)



(i) Llama-3.1-8B-Instruct (MMLU-Pro-Law-100Q)



(j) Phi-3.5-mini-instruct (MMLU-Pro-Law-100Q)

Figure 8: Layer-wise Spearman correlation coefficients between format edit distance and embedding distance across tasks.

F Correlation between Format Edit Distance and Output Consistency

Task	Model	Zero-shot	Zero-shot CoT	Few-shot	Few-shot CoT
CommonsenseQA	Phi-3.5-mini-instruct	-0.74 (6.276e-06)	-0.61 (5.079e-04)	-0.61 (5.343e-04)	-
	Phi-3.5-vision-instruct	-0.70 (3.208e-05)	-0.72 (1.868e-05)	-0.67 (1.047e-04)	-
	Llama-3.1-8B	-	-	-0.69 (5.278e-05)	-
	Llama-3.1-8B-Instruct	-0.50 (0.007)	-0.60 (6.742e-04)	-0.20 (0.309)	-
	Llama-3.1-70B-Instruct	-0.47 (0.011)	-0.62 (3.832e-04)	-0.26 (0.187)	-
	DeepSeek-R1-Distill-Llama-8B	-	-0.60 (6.771e-04)	-	-
	gpt-4o-2024-11-20	-0.46 (0.015)	-0.36 (0.062)	-0.33 (0.082)	-
QASC	Phi-3.5-mini-instruct	-0.57 (0.001)	-0.48 (0.009)	-0.68 (6.519e-05)	-0.48 (0.010)
	Phi-3.5-vision-instruct	-0.71 (1.921e-05)	-0.68 (5.944e-05)	-0.61 (5.923e-04)	-0.72 (1.628e-05)
	Llama-3.1-8B	-	-	-0.61 (4.966e-04)	-0.53 (0.004)
	Llama-3.1-8B-Instruct	-0.49 (0.009)	-0.69 (4.883e-05)	-0.21 (0.283)	-0.20 (0.309)
	Llama-3.1-70B-Instruct	-0.37 (0.053)	-0.38 (0.043)	-0.19 (0.327)	-0.66 (1.463e-04)
	DeepSeek-R1-Distill-Llama-8B	-	-0.59 (9.278e-04)	-	-0.53 (0.003)
	gpt-4o-2024-11-20	-0.58 (0.001)	-0.22 (0.251)	-0.41 (0.031)	-0.20 (0.311)
100TFQA	Phi-3.5-mini-instruct	-0.54 (0.003)	-0.16 (0.404)	-0.40 (0.037)	-0.23 (0.243)
	Phi-3.5-vision-instruct	-0.69 (4.913e-05)	-0.27 (0.169)	-0.48 (0.009)	-0.43 (0.021)
	Llama-3.1-8B	-	-	-0.51 (0.006)	-0.43 (0.022)
	Llama-3.1-8B-Instruct	-0.32 (0.101)	-0.16 (0.423)	-0.49 (0.008)	-0.40 (0.033)
	Llama-3.1-70B-Instruct	-0.19 (0.342)	-0.34 (0.076)	-0.37 (0.052)	-0.22 (0.253)
	DeepSeek-R1-Distill-Llama-8B	-	-0.15 (0.461)	-	-0.16 (0.405)
	gpt-4o-2024-11-20	-0.20 (0.309)	-0.12 (0.536)	-	-0.16 (0.430)
GSM8K	Phi-3.5-mini-instruct	-	-0.43 (0.021)	-	-0.80 (3.629e-07)
	Phi-3.5-vision-instruct	-	-0.39 (0.041)	-	-0.87 (2.870e-09)
	Llama-3.1-8B	-	-	-	-0.66 (1.406e-04)
	Llama-3.1-8B-Instruct	-	-0.59 (9.793e-04)	-	-0.40 (0.036)
	Llama-3.1-70B-Instruct	-	-0.49 (0.008)	-	-0.54 (0.003)
	DeepSeek-R1-Distill-Llama-8B	-	-0.74 (6.255e-06)	-	-0.59 (8.564e-04)
	gpt-4o-2024-11-20	-	-0.23 (0.230)	-	-0.12 (0.527)
MMLU-Pro-Law-100Q	Phi-3.5-mini-instruct	-0.46 (0.015)	-0.40 (0.034)	-0.43 (0.023)	-0.51 (0.006)
	Phi-3.5-vision-instruct	-0.46 (0.013)	-0.48 (0.009)	-0.49 (0.008)	-0.56 (0.002)
	Llama-3.1-8B	-	-	-0.51 (0.006)	-0.42 (0.028)
	Llama-3.1-8B-Instruct	-0.50 (0.007)	-0.45 (0.016)	-0.42 (0.025)	-0.54 (0.003)
	Llama-3.1-70B-Instruct	-0.06 (0.780)	-0.30 (0.123)	-0.29 (0.129)	-0.27 (0.159)
	DeepSeek-R1-Distill-Llama-8B	-	-0.41 (0.028)	-	-0.53 (0.004)
	gpt-4o-2024-11-20	-0.20 (0.297)	-0.11 (0.587)	-0.27 (0.170)	-0.17 (0.384)

Table 12: Correlation analysis between format edit distance and pairwise consistency. Each cell represents Spearman correlation coefficient and p-value (in parenthesis).

G Additional Results of Confidence Analysis

Model	Task	Zero-shot	Zero-shot CoT	Few-shot	Few-shot CoT
Llama-3.1-8B-Instruct	CommonsenseQA	0.49 (7.577e-60)	0.60 (8.244e-96)	0.36 (1.804e-30)	-
	QASC	0.46 (8.151e-41)	0.53 (2.562e-55)	0.16 (1.521e-05)	0.19 (6.279e-07)
	100TFQA	0.51 (1.065e-06)	0.54 (1.036e-06)	0.66 (2.269e-11)	0.69 (1.230e-11)
	GSM8K	-	0.67 (6.372e-90)	-	0.24 (1.339e-11)
	MMLU-Pro-Law-100Q	0.59 (1.231e-10)	0.52 (6.745e-08)	0.30 (0.003)	0.04 (0.673)
gpt-4o-2024-11-20	CommonsenseQA	0.46 (1.729e-52)	0.41 (7.530e-41)	0.36 (1.066e-31)	-
	QASC	0.40 (1.084e-29)	0.47 (2.182e-41)	0.36 (9.932e-25)	0.43 (1.149e-34)
	100TFQA	0.44 (5.141e-05)	0.36 (8.912e-04)	-	0.49 (4.179e-06)
	GSM8K	-	0.59 (1.315e-88)	-	0.30 (1.718e-21)
	MMLU-Pro-Law-100Q	0.65 (3.206e-13)	0.49 (2.556e-07)	0.63 (2.759e-12)	0.26 (0.009)

Table 13: Correlation analysis between model confidence and setwise consistency. Each cell represents Spearman correlation coefficient and p-value (in parenthesis).

H Generalization to Extended Format Variations

We have extended the scope of format variations to include more complex formatting elements, such as indentation (e.g., nested inputs with tabbed subclauses) and list structures (e.g., bulleted lists). These extended formats reflect more realistic use cases, where inputs may vary in structural presentation while maintaining semantic equivalence.

In our case study on CommonsenseQA (zero-shot) with 32 format variants, Llama-3.1-8B-Instruct and GPT-4o exhibits higher inconsistencies, 0.18 and 0.15, respectively. For Llama-3.1-8B-Instruct, we confirm strong correlations between semantic consistency and embedding distance (Spearman $r = -0.71$, $p < 0.05$), format distance (Spearman $r = -0.63$, $p < 0.05$) and model confidence (Spearman $r = 0.62$, $p < 0.05$). For GPT-4o, we observe a significant correlation with format distance (Spearman $r = -0.42$, $p < 0.05$) and confidence (Spearman $r = 0.60$, $p < 0.05$). Overall, this extended analysis demonstrates that our findings are generalizable across a wide range of format variations, highlighting the need for more consistent and format-agnostic language models.