

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

Cheongwoong Kang¹, Jongeun Baek¹, Yeonjea Kim¹ and Jaesik Choi^{1,2}

¹KAIST, ²INEEJI



CONTENTS

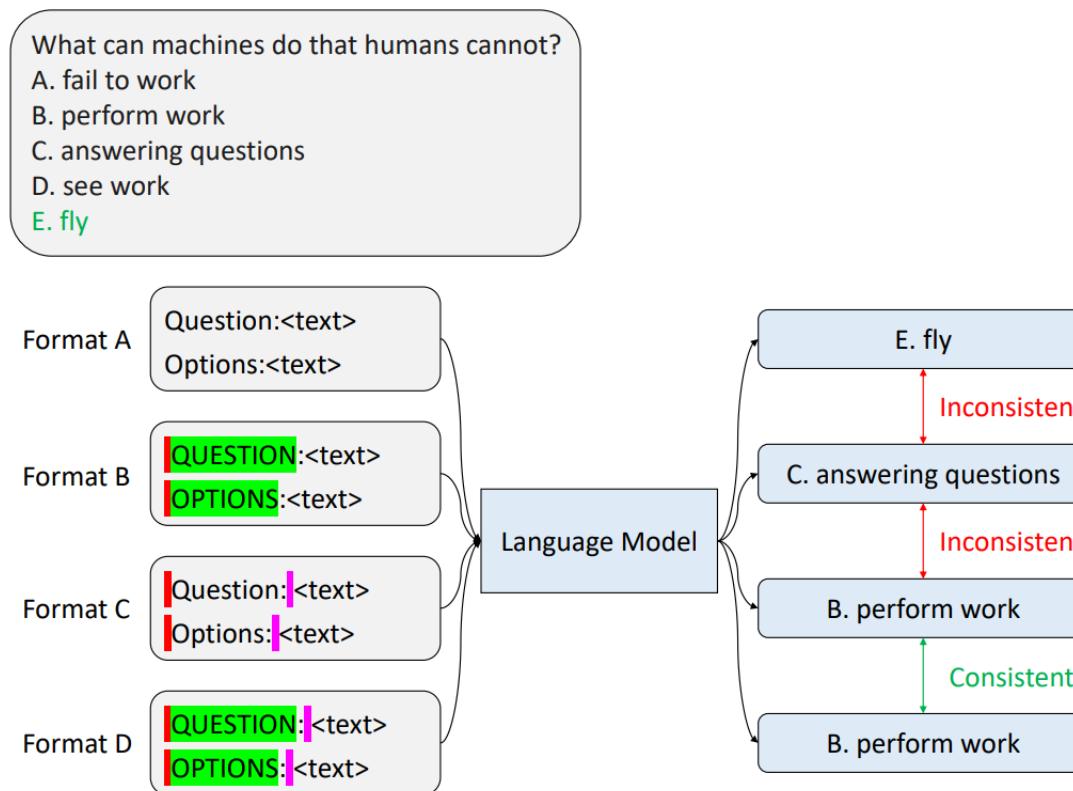
- 01 Introduction
- 02 Quantifying Semantic Inconsistency
- 03 Mechanistic Diagnosis
- 04 Limits of Standard Mitigation Strategies
- 05 Conclusion

01

Introduction

Same Question, Different Answers

- Minor format changes (e.g., spacing, casing) make GPT-4o give different answers to the same question

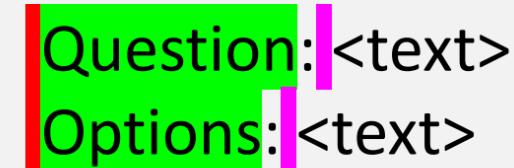


02

Quantifying Semantic Inconsistency

Format Variation Design

- Prompt format variations*: changes in input formatting while preserving the exact wording of the main content
- 3 orthogonal binary factors → $2^3 = 8$ variants per input



Question:<text>
Options:<text>

- Space before descriptor: with or without a space
- Casing of descriptor: capitalized or all caps
- Space after separator: with or without a space

Measuring Consistency

- Pairwise consistency*: measures whether a model returns the same prediction for two semantically equivalent inputs.

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

- Setwise consistency: evaluates whether a model produces the same output across a set of semantically equivalent inputs.

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

Experimental Setup

- A range of datasets, models and prompting strategies

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

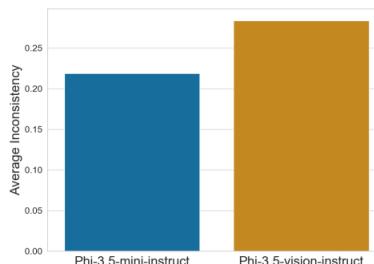
Model	Parameters
Phi-3.5-mini-instruct	3.8B
Phi-3.5-vision-instruct	4.2B
Llama-3.1-8B	8B
Llama-3.1-8B-Instruct	8B
Llama-3.1-70B-Instruct	70B
DeepSeek-R1-Distill-Llama-8B	8B
GPT-4o*	-

* gpt-4o-2024-11-20

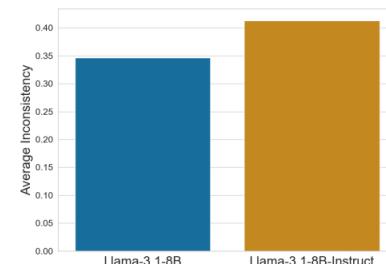
- Prompting strategies
 - Zero-shot: tests pre-trained knowledge
 - CoT (chain-of-thought): encourages intermediate reasoning
 - Few-shot: provides in-context demonstrations to guide behavior

Quantitative Results

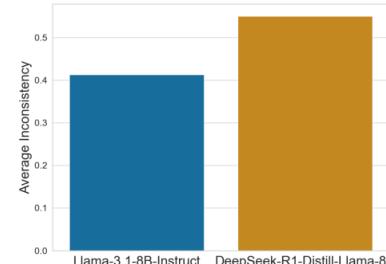
- Takeaways
 - Model scaling reduces inconsistency, while multimodal training, instruction tuning, distillation and prompting strategies show little effect
 - Even GPT-4o remains inconsistent under minor prompt variations
→ **Do LLMs truly separate form from meaning?**



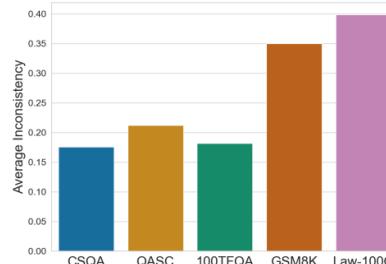
(a) Impact of multimodal training.



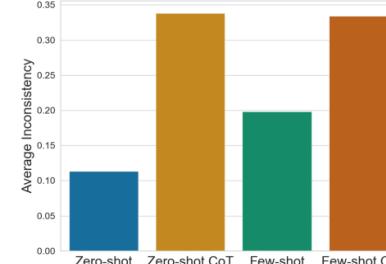
(b) Impact of instruction tuning.



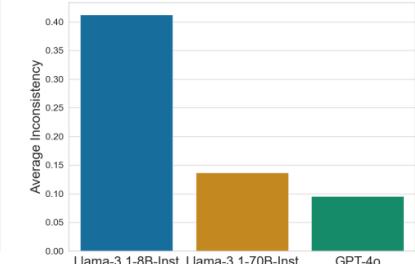
(c) Impact of distillation.



(d) Comparison between tasks.



(e) Comparison between prompting.



(f) Impact of model scaling.

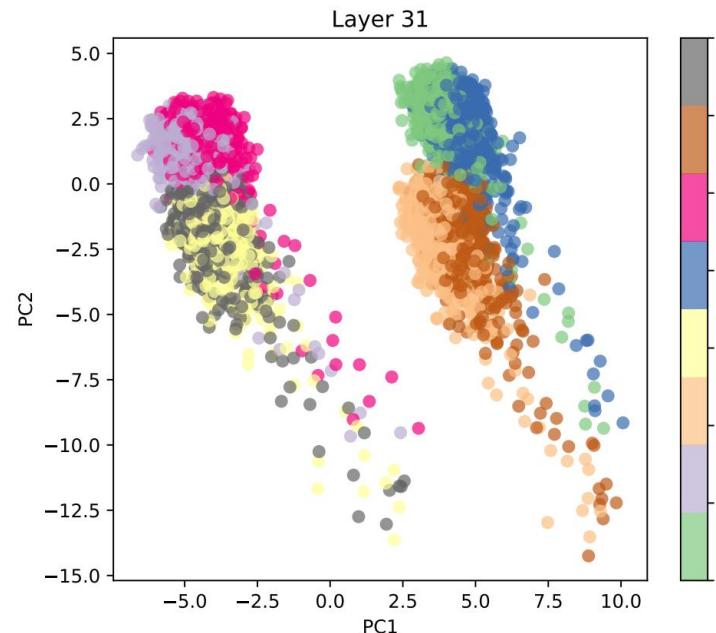


03

Mechanistic Diagnosis

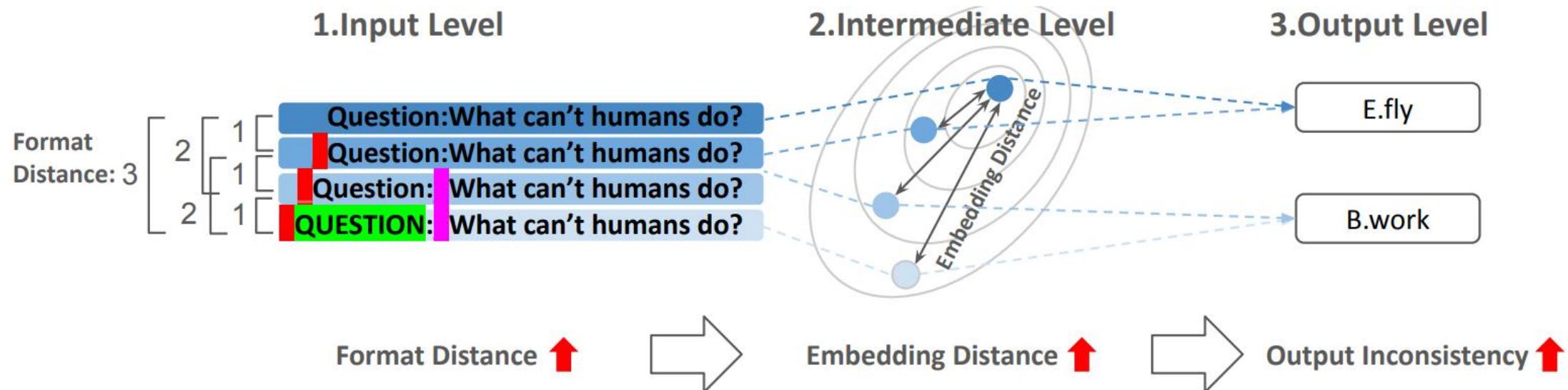
Representation-Level Analysis

- 2D PCA visualization of output embeddings from the final layer at the last token position of the input prompt
 - Distinct formatting variants form separable clusters in the embedding space → format differences are encoded in the model's internal representations



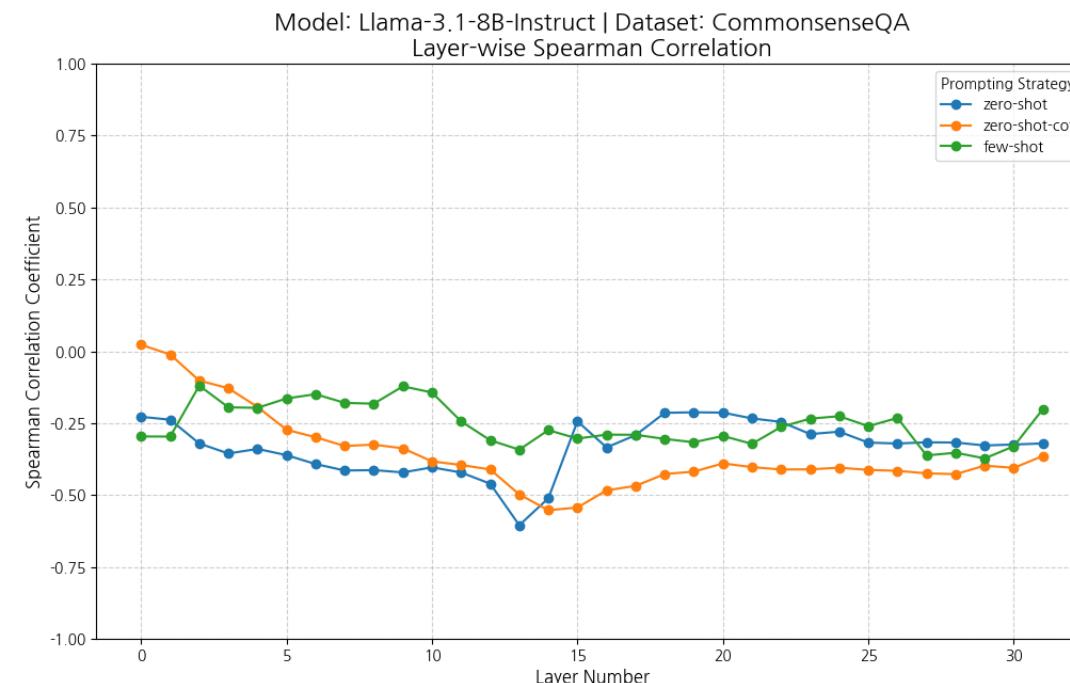
Tracing the Propagation of Inconsistency

- Input-level format changes propagate through embedding space, as embedding distance correlates with output inconsistency → **Form and meaning are entangled** in the embedding space



Layer-wise Correlation Analysis

- Middle layers show strongest link between embedding distance and inconsistency → core semantic representation is most vulnerable to format noise





04

Limits of Standard Mitigation Strategies

Testing Mitigation Strategies

- Prompt-based
 - Reference-guide^{*}: improves consistency, requires external evidence
 - Self-consistency^{**}: improves CoT consistency, computationally expensive

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	-
QASC	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	-
	Llama-3.1-8B-Instruct	0.82 ± 0.09	0.82 ± 0.19	0.61 ± 0.84	0.68 ± 0.74
100TFQA	+ 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
GSM8K	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	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
	Llama-3.1-8B-Instruct	-	0.54 ± 0.75	-	0.79 ± 0.37
GPT-4o	+ 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
GPT-4o	+ Self-consistency	-	0.94 ± 0.08	-	0.96 ± 0.03

^{*}(Lewis et al., 2020)

^{**}(Wang et al., 2022)

Testing Mitigation Strategies

- Fine-tuning on diverse formats
 - Poor generalization to unseen formats: no improvement even for memorized samples when prompted in unseen formats (*surprising failure*)
 - Limited performance gain: some improvement for unseen samples in seen formats, but inconsistency remains



05 Conclusion

Conclusion

- Semantic inconsistency is a representational failure
 - Form and meaning are entangled in embedding space
 - Future solutions must target representation-level alignment

Thank You