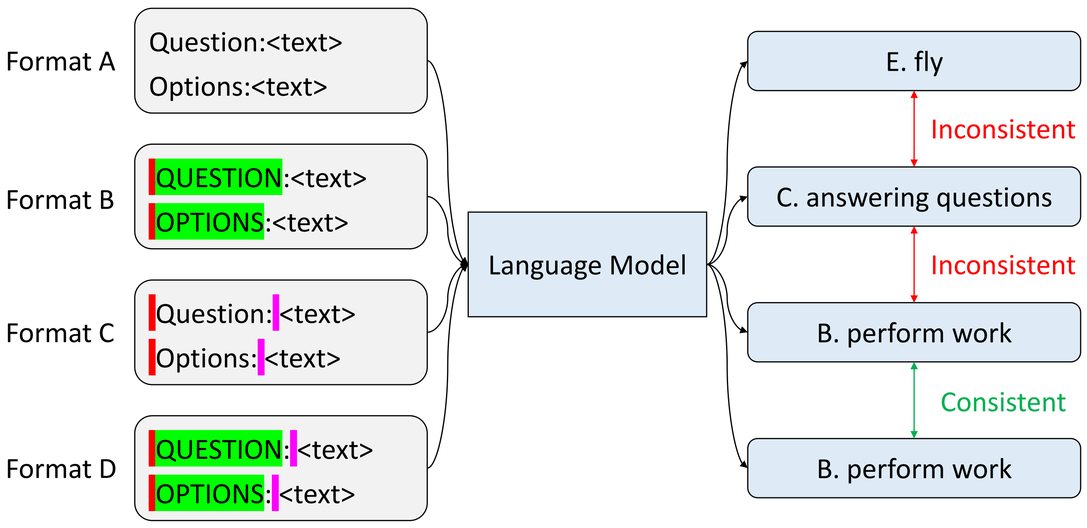
When Format Changes Meaning: Investigating Semantic Inconsistency of Large Language Models

Cheongwoong Kang[1], Jongeun Baek[1], Yeonjea Kim[1], Jaesik Choi[1,2]

[1]KAIST [2]INEEJI

Same Question, Different Answers?

What can machines do that humans cannot?
A. fail to work
B. perform work
C. answering questions
D. see work
E. fly



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

Evaluation Framework

Format Variation Design

- (1) Space before descriptor: with / without
- (2) Casing of descriptor: Capitalized / ALL CAPS
- (3) Space after separator: with / without
- $\Rightarrow 2^3 = 8$ format variants per input

Metrics

- **Semantic consistency**: identical predictions across semantically equivalent inputs

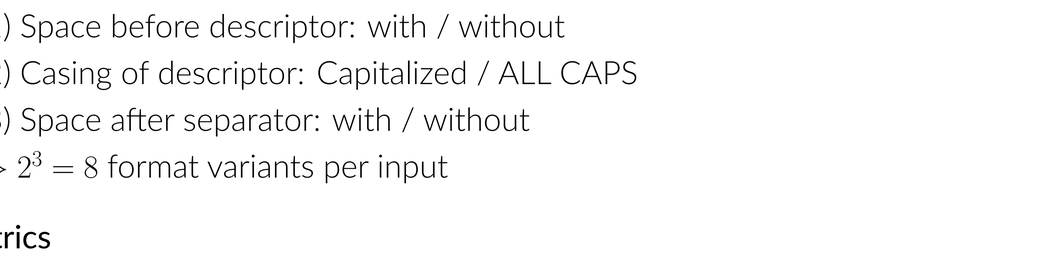
$$\text{Consistency}_{\text{pairwise}}(j,k) = 1(\hat{y}_j = \hat{y}_k) \quad (1)$$

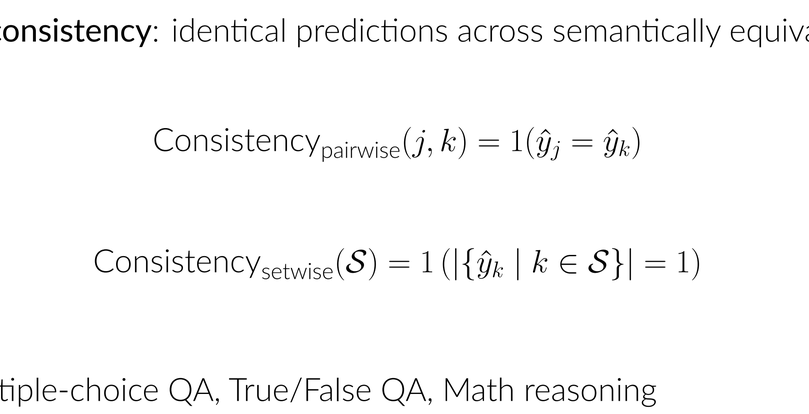
$$\text{Consistency}_{\text{setwise}}(\mathcal{S}) = 1\left(|\{\hat{y}_k \mid k \in \mathcal{S}\}| = 1\right) \quad (2)$$

Setup

- **Tasks**: Multiple-choice QA, True/False QA, Math reasoning
- **Models**: GPT-4o; Llama-3.1 (8B-base, 8B-instruct, 70B-instruct); Phi-3.5-instruct (mini, vision); DeepSeek-R1 (distilled to Llama-3.1-8B-base)
- **Prompting**: Zero-shot, Few-shot, Zero-shot CoT, Few-shot CoT

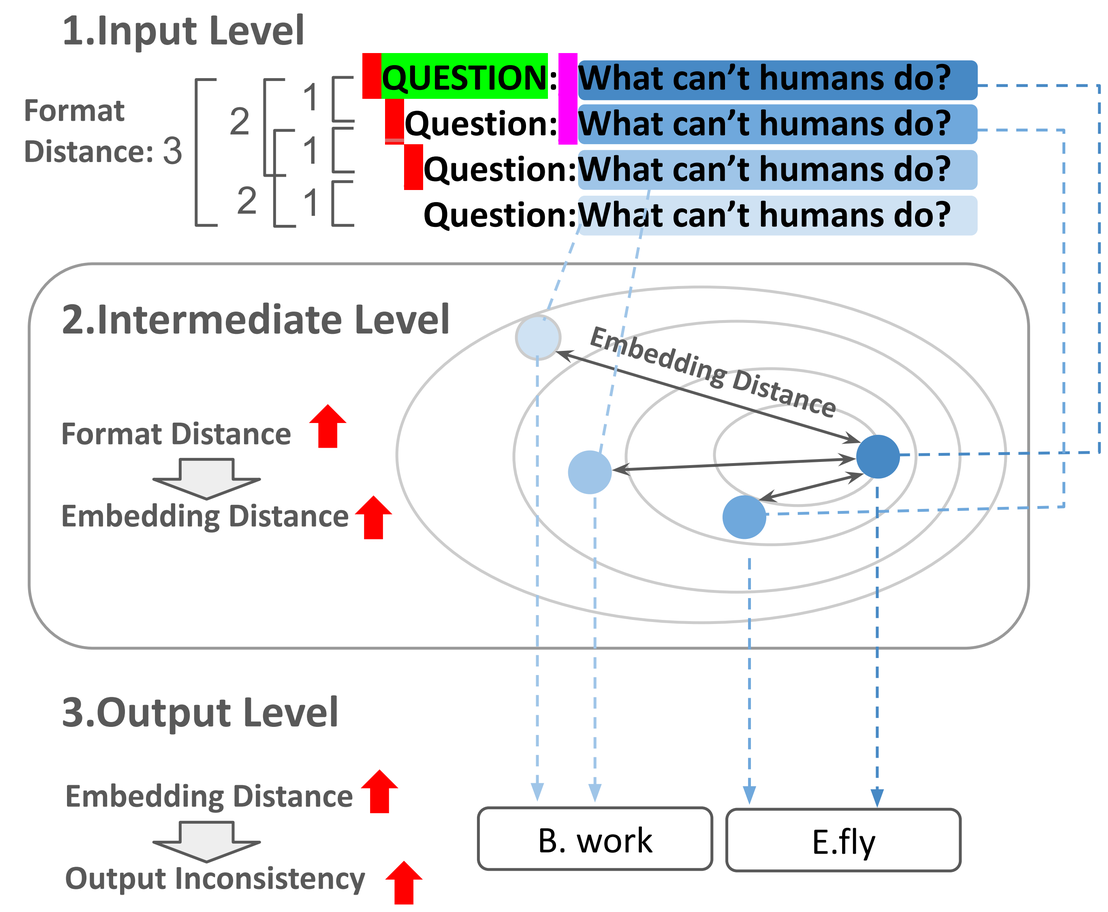
Quantitative Results



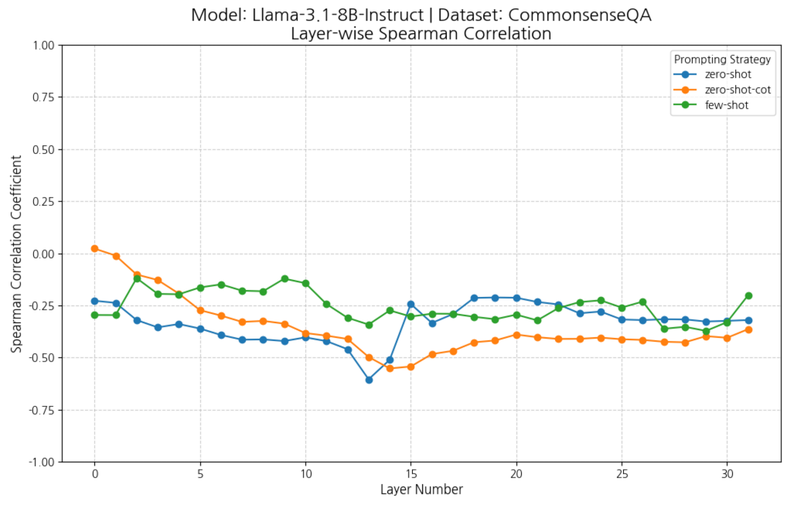


- Greater inconsistency in complex, realistic tasks (e.g., Math, Law)
- **Model scaling** reduces inconsistency, while multimodal training, instruction tuning, distillation and **prompting strategies** show little effect.
- Yet, even GPT-4o remains inconsistent under minor format variations.
  => **Do LLMs truly separate form from meaning?**

Representation-Level Analysis



Embedding instability explains inconsistency: Input-level format changes propagate through embeddings, causing output inconsistency. => Form and meaning are entangled in the embedding space, as format shifts induce semantic drift.



Layer-wise effect: Middle layers show the strongest link between embedding distance and inconsistency, implying that core semantic representation is most vulnerable to format noise.

Context-sensitive impact: Formatting elements (spacing, casing, separator) have similar average effects, but their impact is highly context-dependent, indicating that no single element consistently drives inconsistency.

Uncertainty as a signal: Higher model confidence correlates with higher consistency, suggesting that uncertainty could be leveraged to detect or mitigate inconsistency.

Mitigation Strategies

Prompt-based

- Reference-guided: V improves consistency, X requires external evidence
- Self-consistency: V improves CoT consistency, X computationally expensive

Fine-tuning on diverse formats

- X **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

Summary: Existing methods offer limited improvements. Semantic inconsistency is deeper than input-output mismatch.

Conclusion

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

GitHub: https://github.com/cheongwoong/rolm

Email: cw.kang@kaist.ac.kr