Do LLMs estimate uncertainty well in instruction-following?

Juyeon Heo[1] , Miao Xiong[2] , Christina Heinze-Deml[3] , and Jaya Narain[3]
University of Cambridge[1], National University of Singapore[2], Apple[3]

Motivation

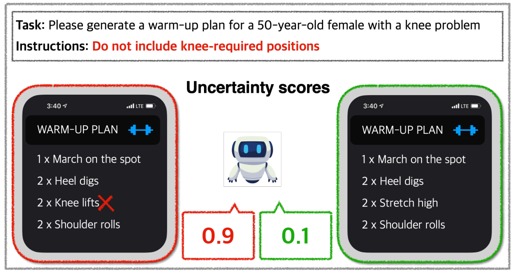
Instruction-following matters for building reliable LLM agent.

Deployed models must strictly follow the instructions and constraints from users to ensure that the outputs are both safe and aligned with user intentions.

Since LLMs are prone to errors, uncertainty estimation ability in instruction-following is essential.

If the LLM misinterprets or deviates from these instructions but accurately recognizes and signals high uncertainty, it could prompt further review or intervention, thereby preventing the delivery of potentially harmful advice.

Example of personalized AI agent: psychological counseling

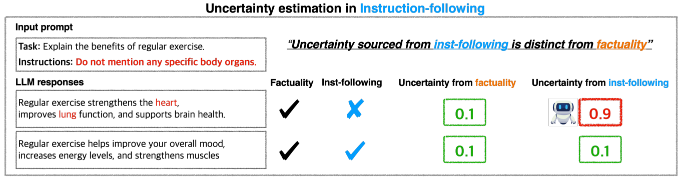


Task: Please generate a warm-up plan for a 50-year-old female with a knee problem
Instructions: Do not include knee-required positions

However, uncertainty estimation in instruction following tasks has received limited attention.

Instruction-following tasks focus on whether a model's response adheres to a set of given instructions, rather than estimating the factual accuracy (Figure 1). Given these different source of uncertainty, it is unclear whether existing methods, which are typically designed for estimating factual uncertainty, can accurately capture uncertainty in instruction following.

Uncertainty estimation in Instruction-following



Input prompt

Task: Explain the benefits of regular exercise.
Instructions: Do not mention any specific body organs.

"Uncertainty sourced from inst-following is distinct from factuality"

LLM responses	Factuality	Inst-following	Uncertainty from factuality	Uncertainty from inst-following
Regular exercise strengthens the heart, improves lung function, and supports brain health.	✓	✗	0.1	0.9
Regular exercise helps improve your overall mood, increases energy levels, and strengthens muscles	✓	✓	0.1	0.1

Methods

Let's evaluate uncertainty estimation in instruction-following tasks using existing datasets.

Uncertainty estimation methods
We employ three types;
Self-evaluation of their own uncertainty
- verbalized confidence(Lin et al., 2022),
- p(true) (Kadavath et al., 2022),

Logits-based method
- perplexity,
- sequence probability,
- mean token entropy (Fomicheva et al., 2020),

Linear probe (Liu et al., 2024)

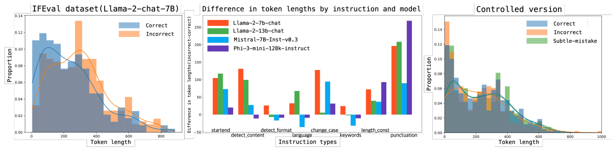
Data We use IFEval dataset (Zhou et al., 2023), which is designed to evaluate the instruction-following ability of LLMs on 25 verifiable instruction types.

Models We evaluate four LLMs of varying sizes:
LLaMA2-chat-7B (Touvron et al., 2023),
LLaMA2-chat-13B (Touvron et al., 2023),
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023),
Phi-3-mini-128k-instruct (Abdin et al., 2024).

However, using existing dataset has limitations.

For example, uncertainty sourced from task execution quality is entangled with uncertainty stemming from instruction-following. Furthermore, there exists length bias, incorrect responses tend to be longer than correct ones across different LLMs in the IFEval dataset, thus models are only evaluated in length-biased settings, missing comparisons on controlled, length-neutral conditions

Case	Task quality entanglement		
	Inst-following	Task quality	Verbalized confidence
Case 1	o	H	7.53 ± 0.49
Case 2	o	L	7.00 ± 0.63
Case 3	x	H	7.42 ± 0.51
Case 4	x	L	6.64 ± 0.33



(a) Length distribution on IFEval (b) Length differences by instruction types (c) Length in Controlled version

Thus, we suggest new benchmark datasets.

To disentangle the complexities that can obscure uncertainty estimation, we design two distinct versions of the dataset: Controlled and Realistic. The Controlled version neutralizes the influence of token length. Meanwhile, the Realistic version leverages actual LLM-generated responses that naturally incorporate real-world signals, including length signal.

Results

Our key findings

Verbalized method consistently outperforms logit-based methods like perplexity in the Controlled-Easy setting, where correct and incorrect responses are relatively easier to distinguish. Specifically, normalized p(true) (Kadavath et al., 2022) proves to be a reliable uncertainty method across both Controlled-Easy and Realistic settings.

Smaller models often outperform larger ones in verbalized confidence, suggesting that factors beyond model size, such as tuning or architecture, may contribute to better uncertainty estimation in certain tasks.

Probes relying on the internal states of LLMs outperform logit-based and verbalized confidence, highlighting promising directions for future.

All approaches, including internal representations, struggle in challenging tasks like Controlled-Hard, which involve subtle off-target responses, to estimate uncertainty accurately, pointing to inherent limitations in LLMs' ability to handle complex uncertainty.

Contribution and Future work

Contributions: We present the first systematic evaluation of uncertainty estimation methods in instruction-following tasks, addressing a gap in existing.

Limitations: the narrow scope of instruction types and domains included in the benchmark, and a potential risk of leakage affecting the evaluation.

Future work: expanding the benchmark to include a broader range of domains and evaluating more LLMs would further deepen understanding of uncertainty estimation in instruction-following. Additional analysis could also investigate why LLMs tend to fail to provide accurate uncertainty estimates in instruction-following, which could lead to the development of trustworthy AI agents.