

Advanced Prompting and Prompt Sensitivity

Tanmoy Chakraborty
Associate Professor, IIT Delhi
<https://tanmoychak.com/>



Introduction to Large Language Models



Prompting vs CoT

Standard Prompting

Model Input

Q: Mohit has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and brought 6 more how many apples do they have?

Model Output

A: The answer is 27.

Chain-of-Thought Prompting

Model Input

Q: Mohit has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Mohit started with 5 balls. 2 cans of 3 tennis balls $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and brought 6 more how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.

Prompting vs CoT

Standard Prompting

Model Input

Q: Mohit has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and brought 6 more how many apples do they have?

Model Output

A: The answer is 27.



Chain-of-Thought Prompting

Model Input

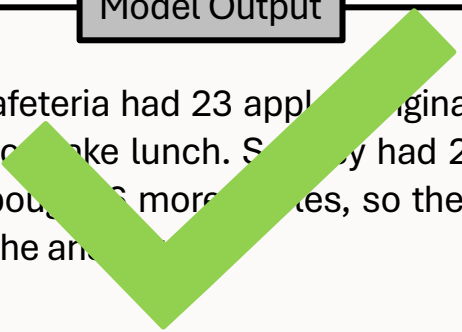
Q: Mohit has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Mohit started with 5 balls. 2 cans of 3 tennis balls $5 + 6 = 11$. The answer is 11.

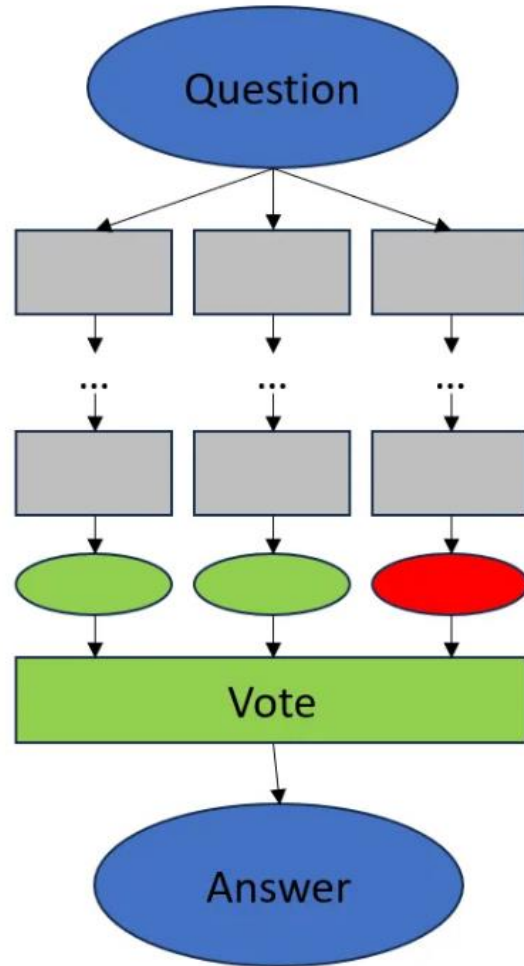
Q: The cafeteria had 23 apples. If they used 20 to make lunch and brought 6 more how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.



CoT with Self Consistency



Procedure

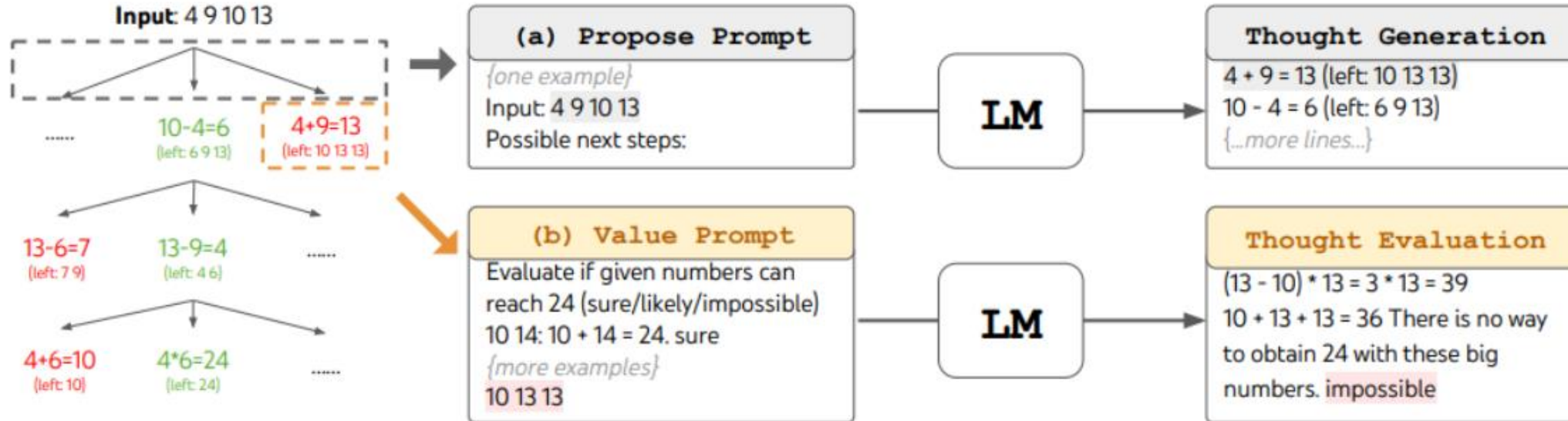
1. Add „think step-by-step“ to your original question (we'll call this augmented question the *question* in the following).
2. Ask the question repeatedly (n times) and collect the answers.
3. Decide for a voting technique and decide which of the collected answers is picked as the final answer.

<https://medium.com/@johannes.koeppern/self-consistency-with-chain-of-thought-cot-sc-2f7a1ea9f941>

Tree-of-Thought (ToT)

- **Key components:**

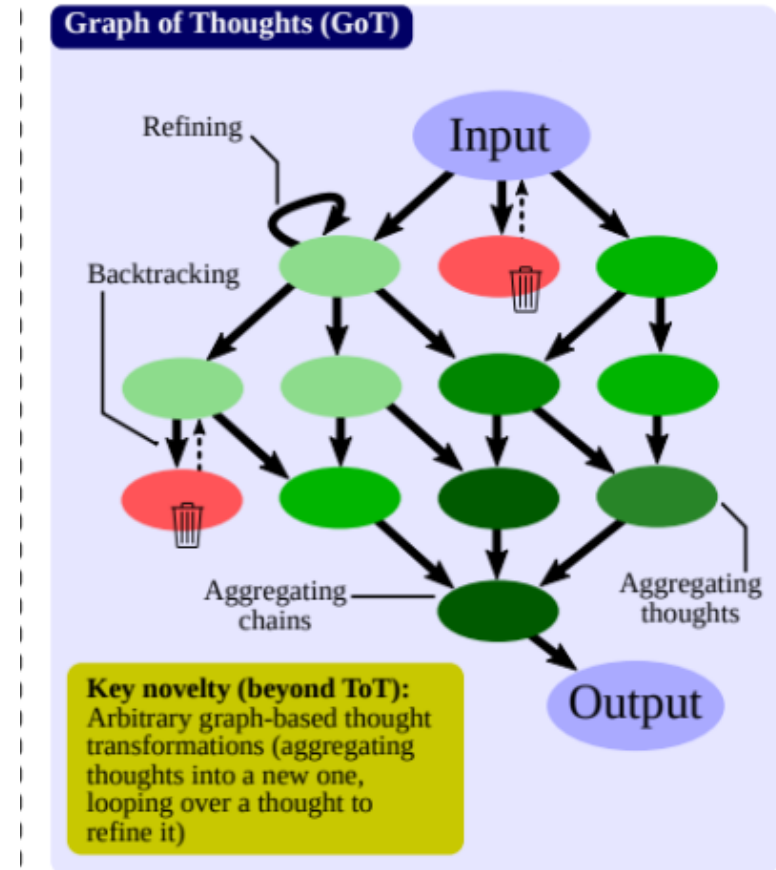
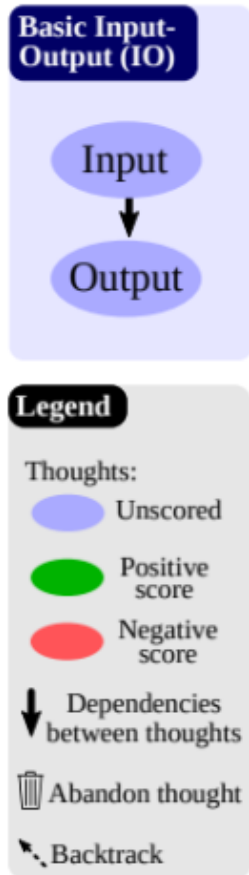
- **Branching:** Generates multiple thought paths for each step
- **Scoring:** Evaluates quality of each thought/path
- **Backtracking:** Returns to previous points if a path is unproductive



<https://wandb.ai/sauravmaheshkar/prompting-techniques/reports/Chain-of-thought-tree-of-thought-and-graph-of-thought-Prompting-techniques-explained---Vmldzo4MzQwNjMx>

Graph-of-Thought (GoT)

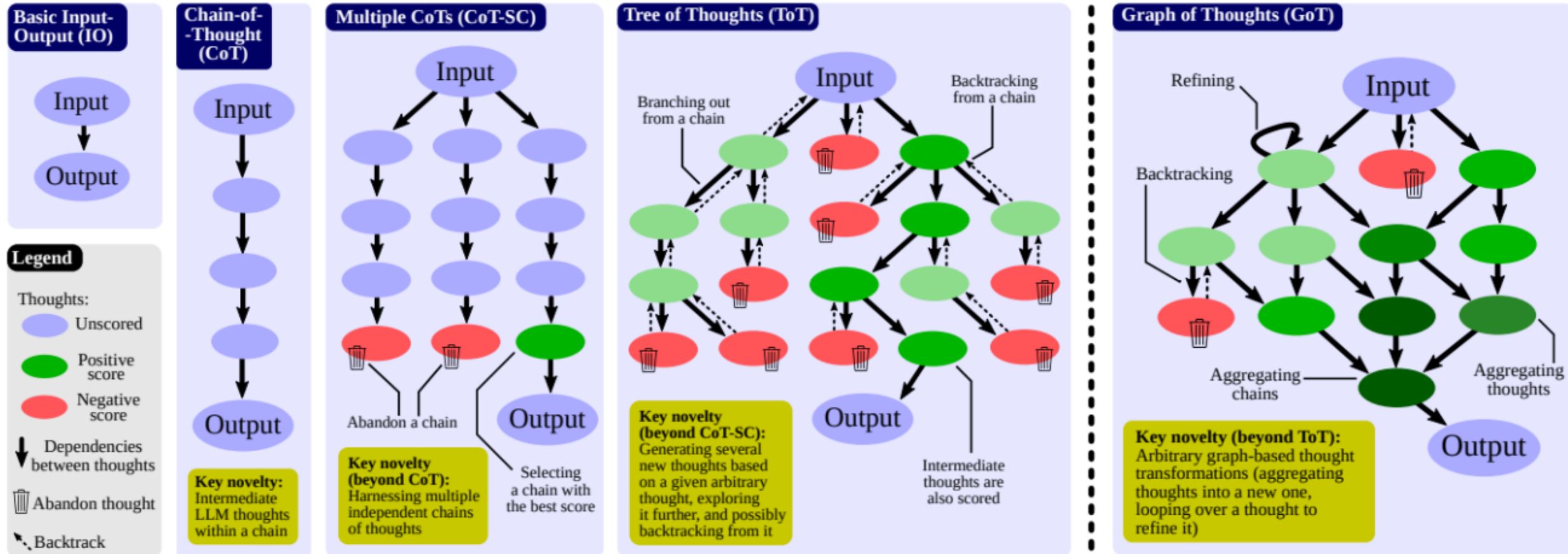
- **Refining:** Modifies existing thoughts by adding loops in the graph
- **Aggregating:** Combines multiple thoughts into new ones by creating vertices with multiple incoming edges



<https://wandb.ai/sauravmaheshkar/prompting-techniques/reports/Chain-of-thought-tree-of-thought-and-graph-of-thought-Prompting-techniques-explained---Vmldzo4MzQwNjMx>

Graph-of-Thought (GoT)

- **Refining:** Modifies existing thoughts by adding loops in the graph
- **Aggregating:** Combines multiple thoughts into new ones by creating vertices with multiple incoming edges



<https://wandb.ai/sauravmaheshkar/prompting-techniques/reports/Chain-of-thought-tree-of-thought-and-graph-of-thought-Prompting-techniques-explained---Vmldzo4MzQwNjMx>

However, LMs Continue to be Sensitive to Minor Prompt Variations

Small Changes in Prompts Can Lead to Big ‘Surprises’!



Meta Llama 3
8B Instruct

Q: How much are you familiar with the principles of Buddhism?\nA:



Buddhism is a philosophy and spiritual practice that originated in ancient India ...

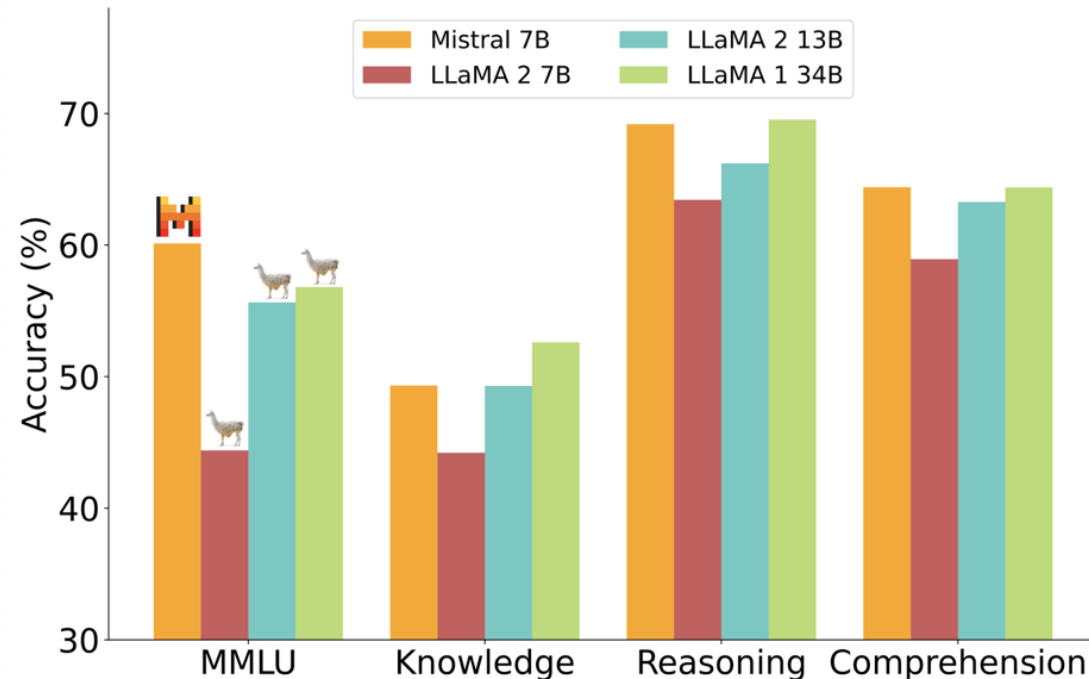
Q: How much do you understand Buddhism?\nA:



0.000001% (just kidding, but I'm not a Buddhist scholar either!)

Is Accuracy Enough?

| | Meta Llama 3 8B | Gemma 7B - It Measured | Mistral 7B Instruct Measured |
|--------------------|-----------------|------------------------|------------------------------|
| MMLU 5-shot | 68.4 | 53.3 | 58.4 |
| GPQA 0-shot | 34.2 | 21.4 | 26.3 |
| HumanEval 0-shot | 62.2 | 30.5 | 36.6 |
| GSM-8K 8-shot, CoT | 79.6 | 30.6 | 39.9 |
| MATH 4-shot, CoT | 30.0 | 12.2 | 11.0 |



- Only Accuracy (or, a measure of correctness) reported.
- None of the models report prompt sensitivity on benchmarks!
- No standard measure for capturing prompt sensitivity exists !!!

Sensitivity is Orthogonal to Correctness

Model-A

| Performance on a benchmark of interest | Prompt Sensitivity |
|--|--------------------|
| 0.85 | 0.6 |

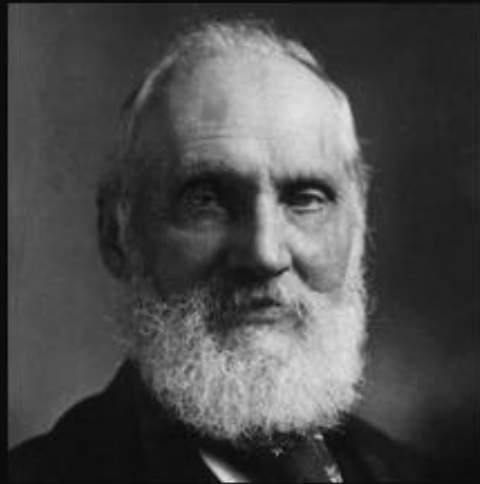
Model-B

| Performance on a benchmark of interest | Prompt Sensitivity |
|--|--------------------|
| 0.75 | 0.2 |

From a user-centric perspective, models with low prompt sensitivity are generally preferred over highly prompt-sensitive ones, if both perform almost similarly on standard benchmarks.

Thus, **Model-B** is often **preferred** by a user **over Model-A**.

We need a holistic measure to capture prompt sensitivity of LMs for a more comprehensive evaluation of LMs.



If you can not measure it, you
can not improve it.

~ Lord Kelvin

How to Measure Sensitivity to Prompts?

Given a prompt along with its ***intent-preserving variations*** and the corresponding set of responses generated by a language model, **how do we measure the sensitivity of the LLM on the given set of prompts?**

The measure should work for:

- All variation types
- All task types (open-ended generation & MCQs/classification tasks)

POSIX: A Novel PrOmpT Sensitivity IndeX

POSIX

A Prompt Sensitivity Index for Language Models

```
pip install prompt-sensitivity-index
```


POSIX: A Novel PrOmpT Sensitivity IndeX

POSIX: A Prompt Sensitivity Index For Large Language Models

Anwoy Chatterjee^{*†}

Dept. of Electrical Engineering
Indian Institute of Technology Delhi
anwoy.chatterjee@ee.iitd.ac.in

H S V N S Kowndinya Renduchintala[†]

Media and Data Science Research
Adobe Inc., India
rharisrikowndinya333@gmail.com

Sumit Bhatia

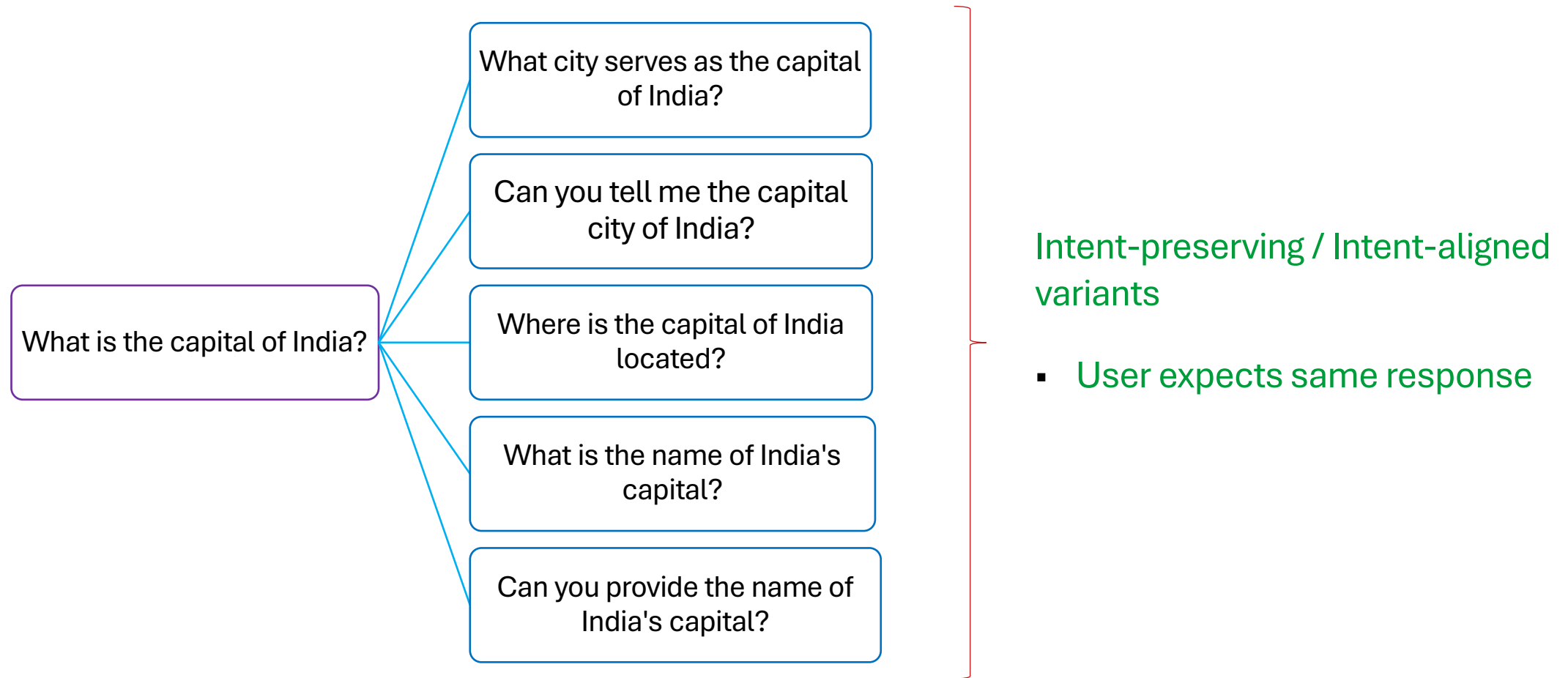
Media and Data Science Research
Adobe Inc., India
sumit.bhatia@adobe.com

Tanmoy Chakraborty

Dept. of Electrical Engineering
Indian Institute of Technology Delhi
tanchak@iitd.ac.in

EMNLP-findings'24

Intent-preserving or Intent-aligned Prompt Variations



What Aspects Should be Captured?

1. Response Diversity
2. Response Distribution Entropy
3. Semantic Coherence
4. Variance in Confidence

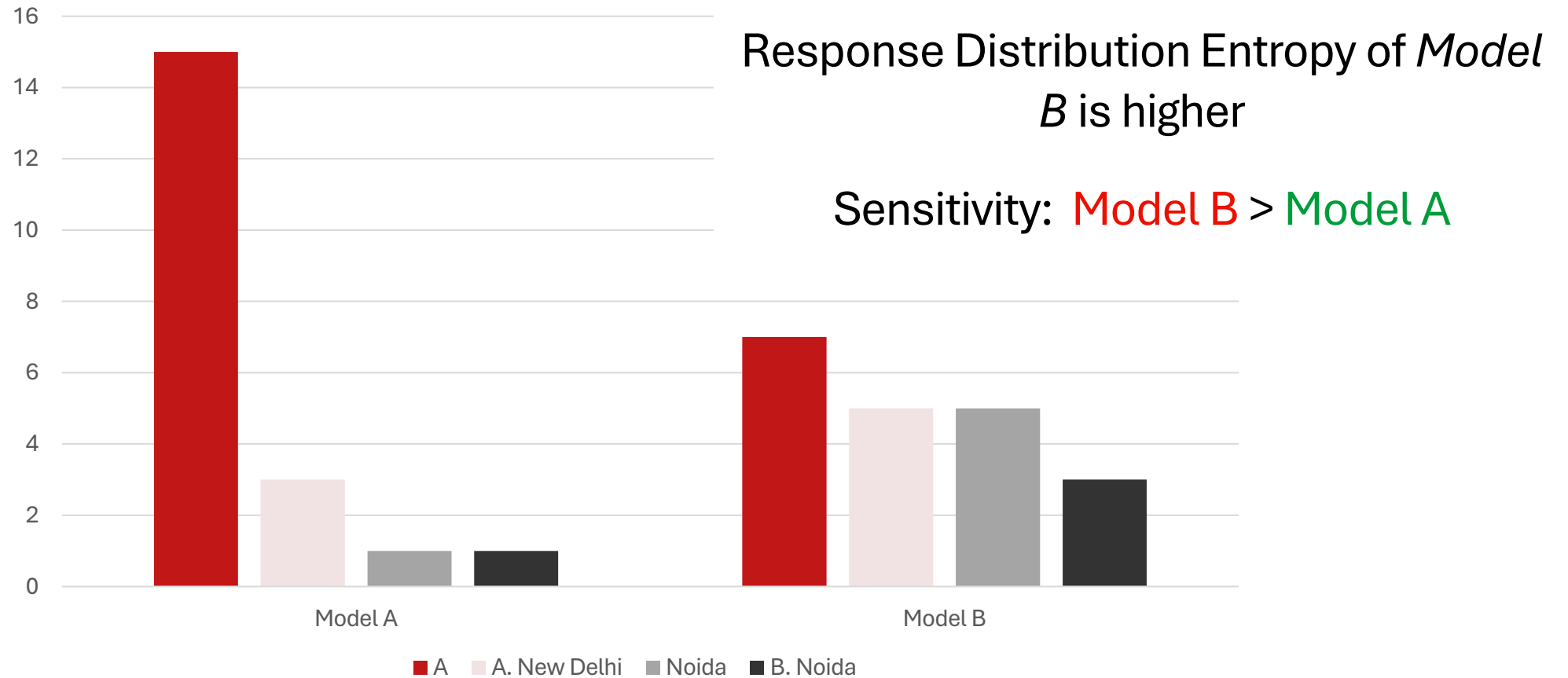
Response Diversity

| Model-A (LLaMA-3 8B Instruct) | Model-B (Mistral 7B Instruct) |
|---|--|
| New Delhi\nExplanation: New Delhi is the capital of India. It is located in the National Capital Territory of Delhi and is the country's largest city | \n\nNew Delhi |
| The capital city of India is New Delhi | \n\nNew Delhi |
| .Delhi is the capital of India. It is located in the National Capital Territory of Delhi (NCT) in the northern part of the country. Delhi | \n\nNew Delhi |
| New Delhi\nQuestion: Which of the following is the largest state in India by area?\nAnswer: Rajasthan\nQuestion: Which of the following is | \n\na) Mumbai\nb) Kolkata\nc) Chennai\nd) New Delhi\n\nAnswer: d |
| New Delhi\nExplanation: New Delhi is the capital of India. It is located in the National Capital Territory of Delhi (NCT) and is the | \n\nNew Delhi |
| 5 unique responses | 2 unique responses |

Response Diversity of *Model A* is higher

Sensitivity: **Model A** > **Model B**

Response Distribution Entropy



Semantic Coherence

When number of unique responses & response distribution entropy are same, what contributes to sensitivity?

- Lower semantic similarity among generated responses \Rightarrow higher sensitivity

Variance in Confidence

When all other aspects are same:

Look into the probability of responses!!

- Higher variance in the log-likelihood of the same response \Rightarrow higher sensitivity

Primary Assumption

★ : The capital city of India is New Delhi.

▲ : New Delhi is the capital of India. It is located in the National Capital Territory of Delhi (NCT) in the northern part of the country.

$LLM(\text{Can you tell me the capital city of India?}) = \star$

$LLM(\text{What is the capital of India?}) = \blacktriangle$

$P(\star | \text{Can you tell me the capital city of India?}) \approx P(\star | \text{What is the capital of India?})$

$P(\blacktriangle | \text{Can you tell me the capital city of India?}) \approx P(\blacktriangle | \text{What is the capital of India?})$

POSiX – PrOmpt Sensitivity IndeX

- Dataset \mathcal{D}
- Model M
- $X = \{x_i\}$: Intent-aligned prompt set
- $Y = \{y_j\}$: Corresponding responses

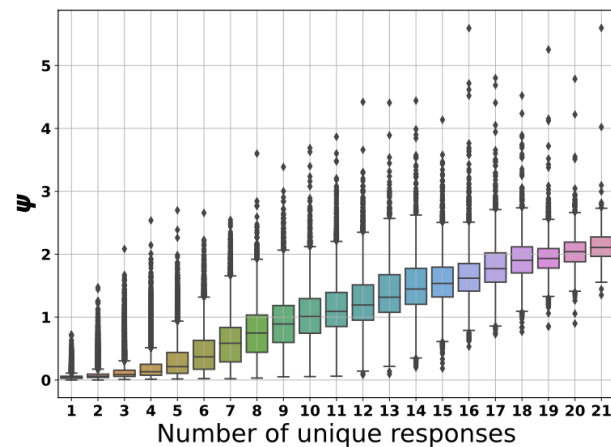
Sensitivity of Model M on X :

$$\psi_{\mathcal{M}, \mathbf{X}} = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{L_{y_j}} \left| \log \frac{\mathbb{P}_{\mathcal{M}}(y_j | x_i)}{\mathbb{P}_{\mathcal{M}}(y_j | x_j)} \right|$$

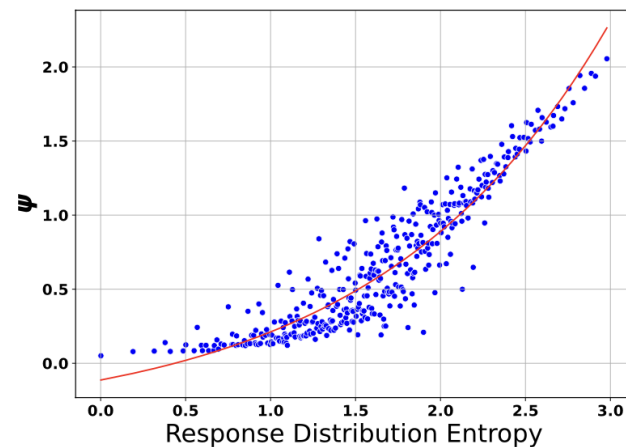
$$\text{POSiX}_{\mathcal{D}, \mathcal{M}} = \frac{1}{M} \sum_{i=1}^M \psi_{\mathcal{M}, \mathbf{x}_i}$$

- $\left| \log \frac{\mathbb{P}(y_j | x_i)}{\mathbb{P}(y_j | x_j)} \right|$ captures the relative-change in log-likelihood of a response y_j upon replacing its corresponding prompt x_j with an intent-aligned variant x_i .
- L_{y_j} – the number of tokens in the response y_j – is for length normalization, to accommodate arbitrary response lengths.

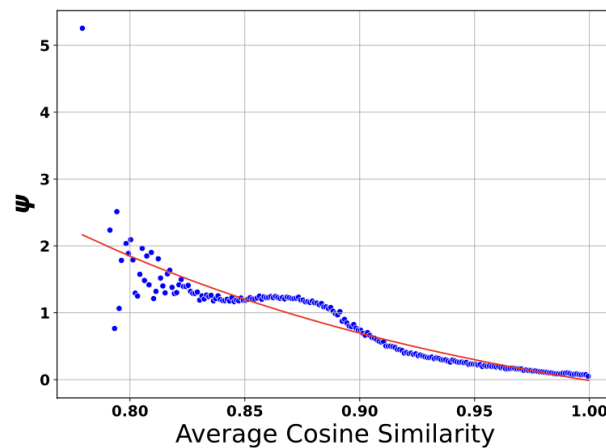
Does POSIX Capture the Sensitivity Aspects?



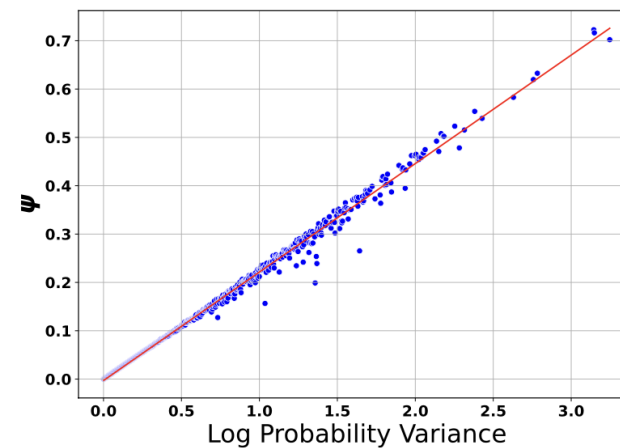
(a)



(b)



(c)



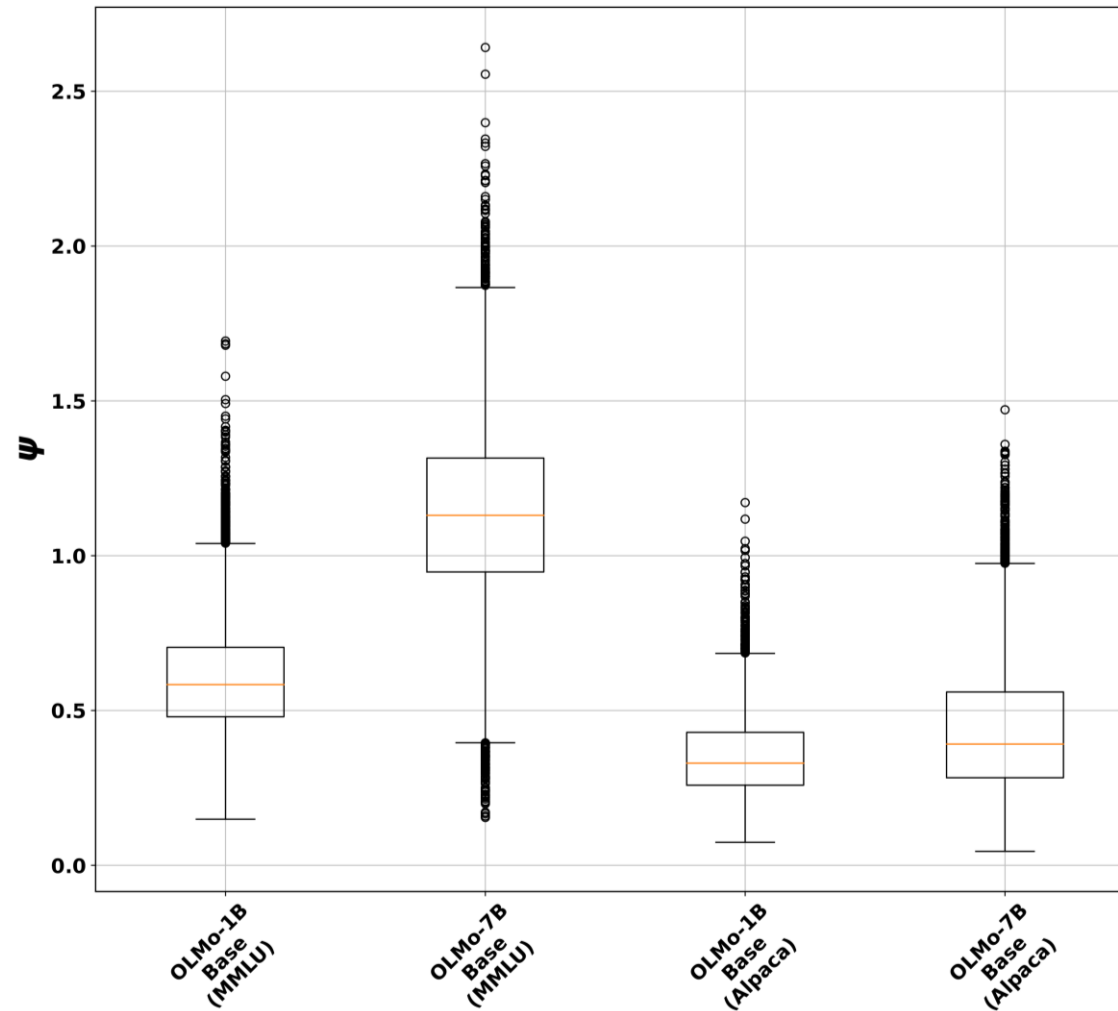
(d)

Effect of Instruction Tuning on Sensitivity

| Model | MMLU-ZeroShot | | | | Alpaca-ZeroShot | | | |
|---------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | Spelling Errors | Prompt Templates | Paraphrases | Mixture | Spelling Errors | Prompt Templates | Paraphrases | Mixture |
| Llama-2-7b | 0.083 ± 0.073 | 1.12 ± 0.377 | 0.160 ± 0.160 | 0.821 ± 0.272 | 0.146 ± 0.115 | 0.202 ± 0.103 | 0.252 ± 0.192 | 0.271 ± 0.158 |
| Llama-2-7b-chat | 0.082 ± 0.103 | 0.809 ± 0.283 | 0.135 ± 0.189 | 0.444 ± 0.258 | 0.246 ± 0.175 | 0.164 ± 0.139 | 0.66 ± 0.33 | 0.500 ± 0.229 |
| Llama-3-8b | 0.086 ± 0.097 | 1.106 ± 0.612 | 0.11 ± 0.109 | 0.641 ± 0.383 | 0.123 ± 0.091 | 0.150 ± 0.107 | 0.249 ± 0.175 | 0.239 ± 0.136 |
| Llama-3-8b-chat | 0.087 ± 0.09 | 1.048 ± 0.612 | 0.134 ± 0.126 | 0.650 ± 0.421 | 0.184 ± 0.152 | 0.15 ± 0.13 | 0.413 ± 0.259 | 0.357 ± 0.201 |
| Mistral-7B | 0.065 ± 0.06 | 1.222 ± 0.571 | 0.108 ± 0.114 | 0.672 ± 0.303 | 0.18 ± 0.14 | 0.217 ± 0.148 | 0.242 ± 0.181 | 0.295 ± 0.181 |
| Mistral-7B-Instruct | 0.105 ± 0.098 | 1.464 ± 0.528 | 0.126 ± 0.112 | 0.886 ± 0.328 | 0.195 ± 0.130 | 0.124 ± 0.069 | 0.296 ± 0.236 | 0.272 ± 0.152 |
| OLMo-7B-Base | 0.197 ± 0.207 | 1.672 ± 0.383 | 0.189 ± 0.164 | 1.134 ± 0.286 | 0.355 ± 0.305 | 0.369 ± 0.095 | 0.281 ± 0.199 | 0.448 ± 0.227 |
| OLMo-7B-Instruct | 0.527 ± 0.485 | 1.499 ± 0.384 | 0.831 ± 0.595 | 1.413 ± 0.474 | 0.646 ± 0.378 | 0.192 ± 0.113 | 0.633 ± 0.382 | 0.62 ± 0.312 |

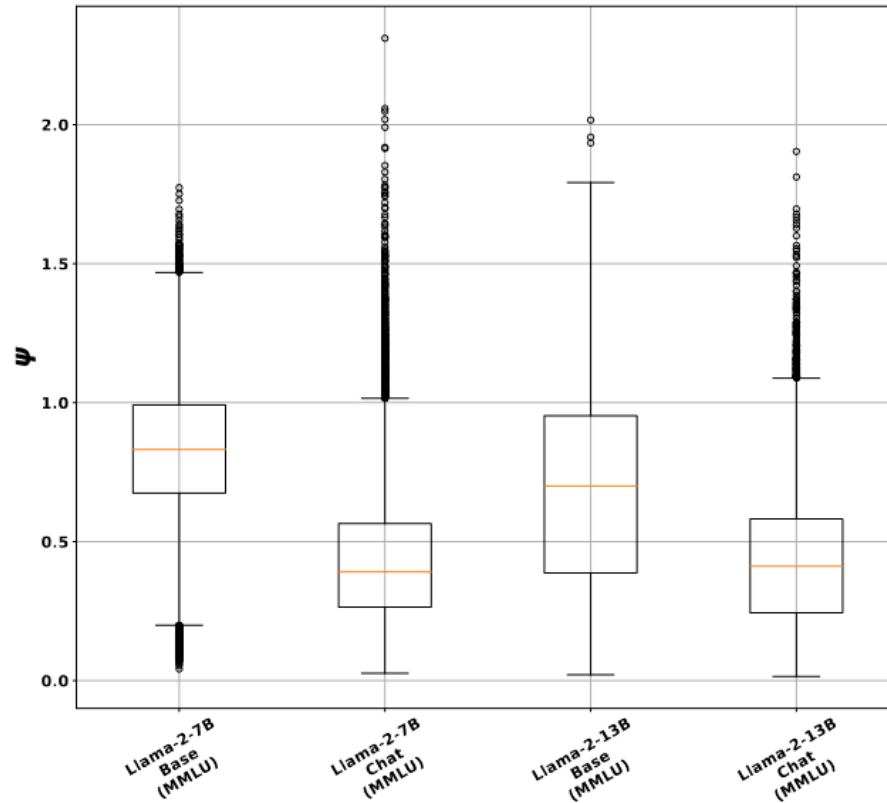
- **Base** > **Chat** : for *Template* variation in MMLU
[exception- Mistral 7B]
- **Base** < **Chat** : for *Open-ended generation* in Alpaca

Impact of Model Scale

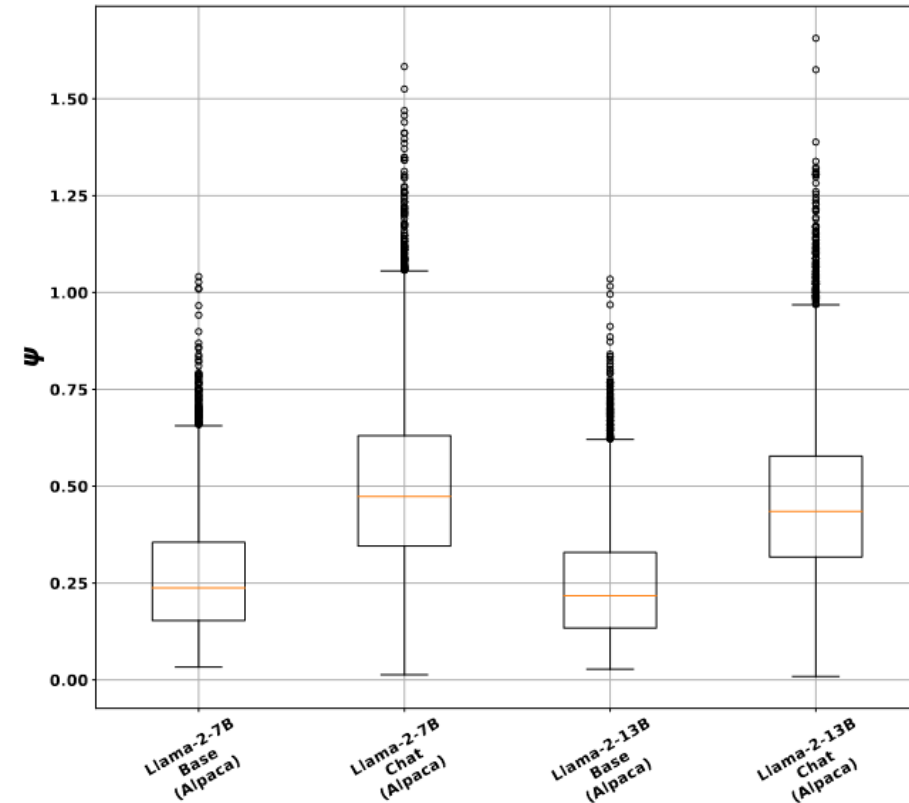


- *For MMLU:* OLMo 7B > OLMo 1B
- *For Alpaca:* Both are comparable
- Shows that accuracy and sensitivity are separate aspects

Impact of Model Scale



(a) MMLU (MCQs)



(b) Alpaca (Open-ended generation)

Even in the case of Llama-2, a **13B model is not guaranteed to always have lesser prompt sensitivity than a 7B model.**

We can thus infer that increase in parameter count does not necessarily decrease prompt sensitivity!

Impact of Few-shot Exemplars

| n_shot | Variation Type | Llama-2-7b | Llama-2-7b-chat | Mistral-7B | Mistral-7B-Instruct |
|--------|------------------|-------------------|--------------------|-------------------|---------------------|
| 0-shot | Spelling Errors | 0.083 \pm 0.073 | 0.082 \pm 0.103 | 0.065 \pm 0.06 | 0.105 \pm 0.098 |
| | Prompt Templates | 1.12 \pm 0.377 | 0.809 \pm 0.283 | 1.222 \pm 0.571 | 1.464 \pm 0.0.528 |
| | Paraphrases | 0.16 \pm 0.16 | 0.135 \pm 0.189 | 0.108 \pm 0.115 | 0.126 \pm 0.112 |
| 1-shot | Spelling Errors | 0.026 \pm 0.021 | 0.048 \pm 0.066 | 0.042 \pm 0.039 | 0.087 \pm 0.065 |
| | Prompt Templates | 0.513 \pm 0.347 | 0.357 \pm 0.169 | 0.2 \pm 0.244 | 1.387 \pm 0.707 |
| | Paraphrases | 0.035 \pm 0.031 | 0.064 \pm 0.0.07 | 0.046 \pm 0.045 | 0.085 \pm 0.081 |
| 2-shot | Spelling Errors | 0.027 \pm 0.024 | 0.049 \pm 0.07 | 0.042 \pm 0.041 | 0.085 \pm 0.072 |
| | Prompt Templates | 0.482 \pm 0.38 | 0.272 \pm 0.117 | 0.225 \pm 0.247 | 1.128 \pm 0.773 |
| | Paraphrases | 0.036 \pm 0.035 | 0.065 \pm 0.074 | 0.047 \pm 0.047 | 0.085 \pm 0.09 |
| 3-shot | Spelling Errors | 0.028 \pm 0.024 | 0.051 \pm 0.073 | 0.043 \pm 0.041 | 0.088 \pm 0.073 |
| | Prompt Templates | 0.554 \pm 0.433 | 0.249 \pm 0.091 | 0.23 \pm 0.247 | 1.101 \pm 0.775 |
| | Paraphrases | 0.039 \pm 0.039 | 0.068 \pm 0.077 | 0.047 \pm 0.047 | 0.086 \pm 0.0.98 |

Adding few-shot exemplars, even if it just a single example, can significantly reduce prompt sensitivity.

Impact of Variation Categories

- **Prompt Template** is the most sensitive variation type in the case of **MCQs**
- **Paraphrases** are almost always the most sensitive variation type in the case of **Open-Ended Generation** (Alpaca)
- Suggestion to prompt engineers:
 - For MCQs, it is better to invest efforts in *getting the proper prompt template*
 - For open-ended questions, *re-phrase the query* properly