Are LLMs Consistent Reasoners?

Authors: Vasilis Karlis, Theofanis Aslanidis, Daniël Van Dijk, Zoë Tzifa-Kratira, Oliver Neut **Supervisors:** Alina Leidinger

1. background

- LLMs’ responses are sensitive to minor textual changes:
  - sycophancy, prompt attacks, typos…
- 🤔 can LLMs consistently reason?
- ➡️ Need for systematic evaluations

- NLI is a representative task for assessing the logical reasoning ability of LLMs
  - LoNLI: Logical natural language inference
- Chain of thought:

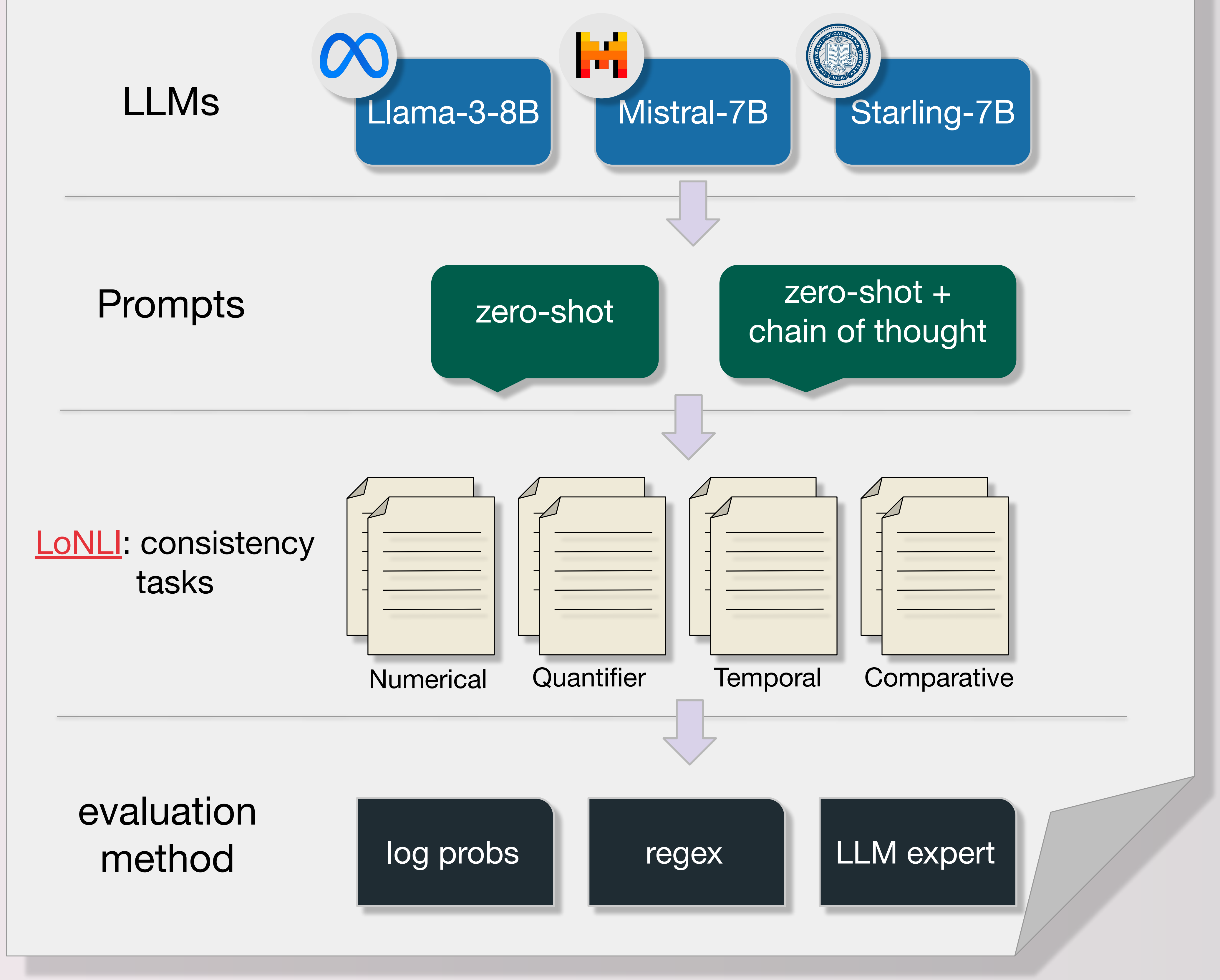
Q: “Paul has 5 dollars. He received 6 more dollars. Paul now has 19 dollars.”
A: “Let’s think step by step”

A: “Let’s think step by step! From the premise, we know: 1. Paul has 5 dollars. 2. He received 6 more dollars.
Now let’s analyze the hypothesis: Paul now has 19 dollars…”

2. research questions:

1. To what extend do LLMs exhibit consistent logical reasoning behavior?
2. Do methods like Chain of Thought enhance consistency and/or performance?
3. How do we evaluate consistency of LLMs on logical reasoning?
4. How do we evaluate performance of LLMs on logical reasoning?

3. methodology



LLMs

Llama-3-8B

Mistral-7B

Starling-7B

Prompts

zero-shot

zero-shot + chain of thought

LoNLI: consistency tasks

Numerical

Quantifier

Temporal

Comparative

evaluation method

log probs

regex

LLM expert

4. results

Log probabilities evaluation

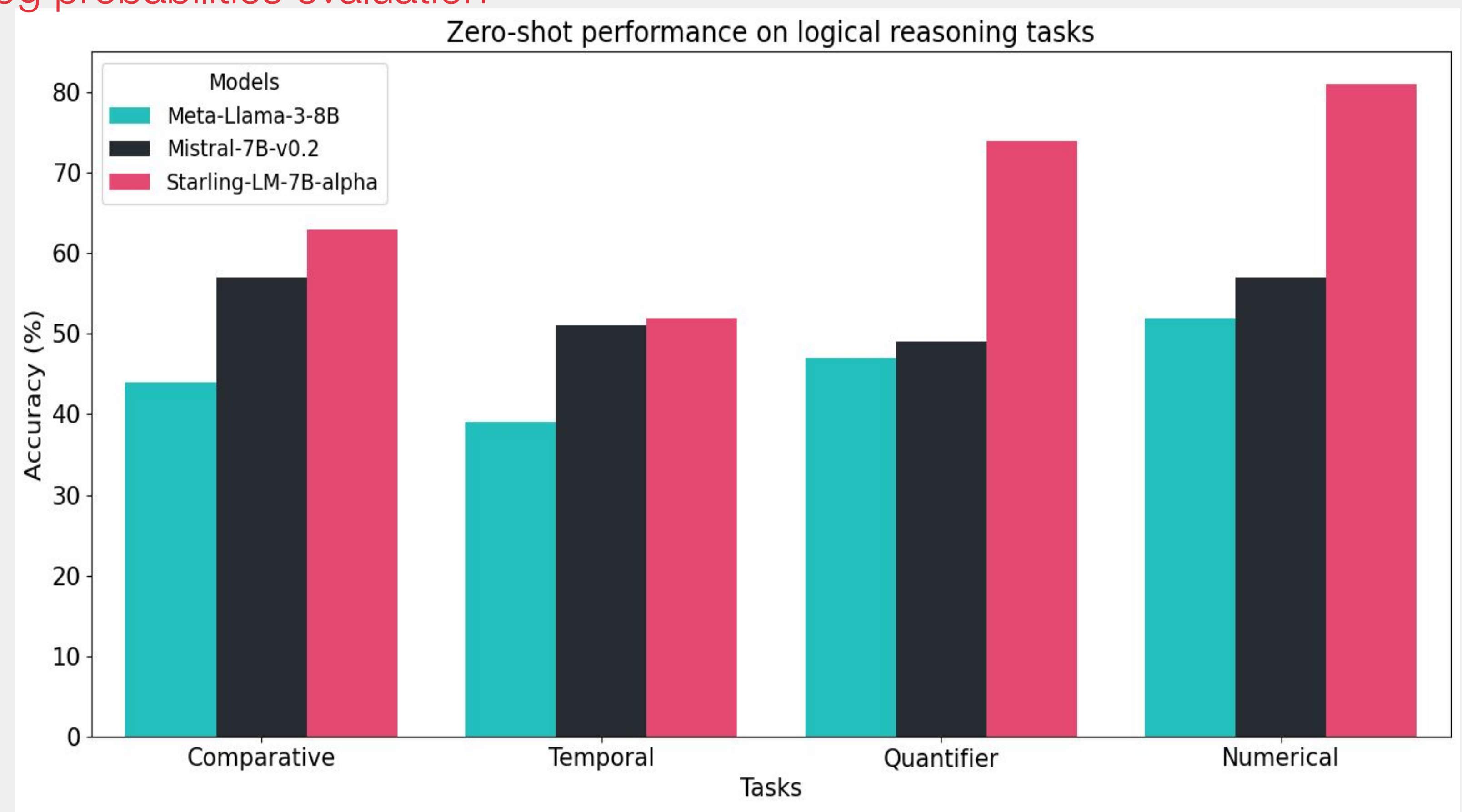


Figure 1: zero-shot performance on logical reasoning tasks using the **log probabilities** evaluation method

4. results

Fine-tuned Mistral 7B evaluator

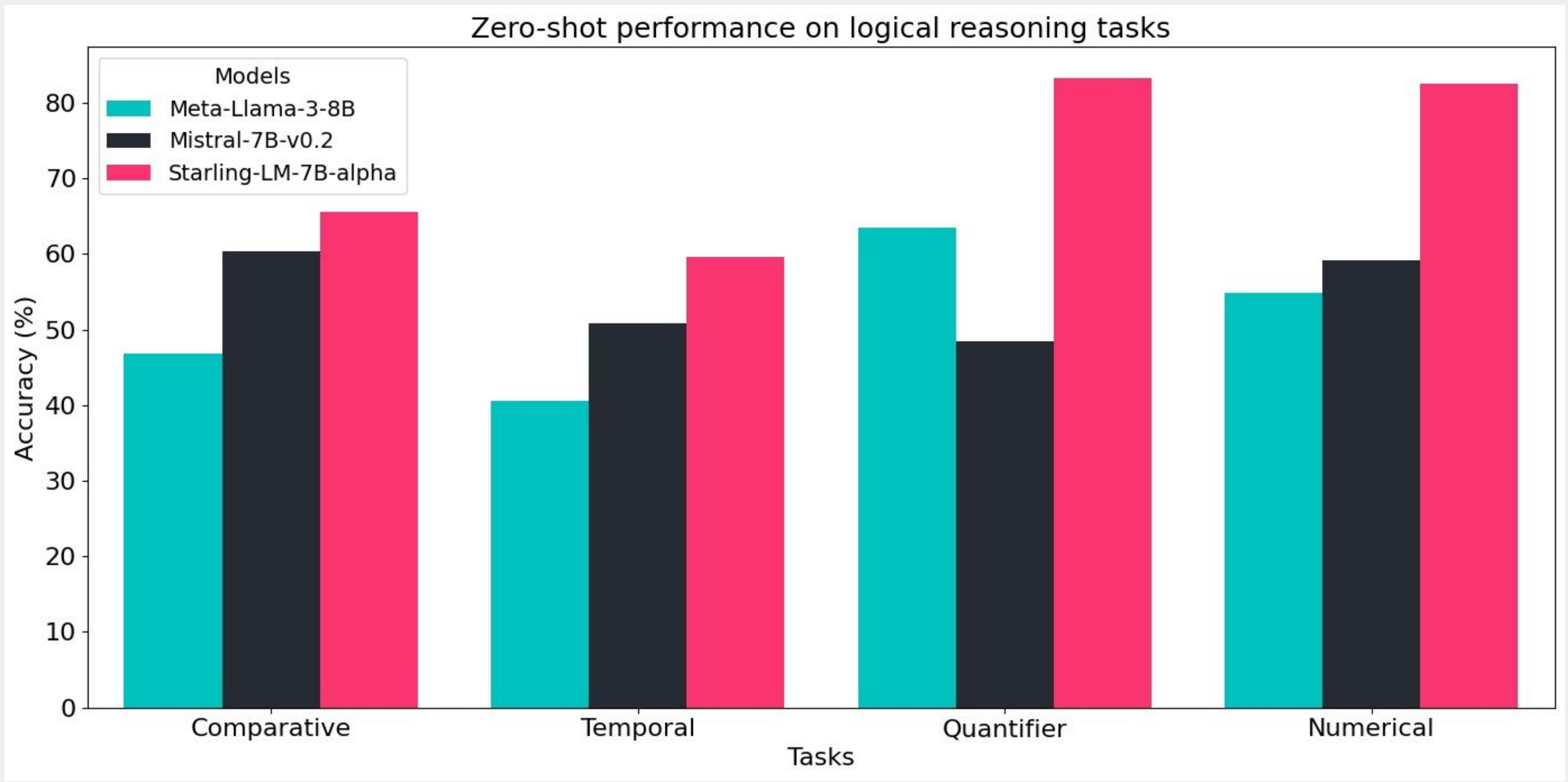


Figure 2: zero-shot performance on logical reasoning tasks using the **fine-tuned LLM evaluator** evaluation method

- Result: Starling-7B is the winner

5. conclusion

1. Even for the same logical task, the same model can have very different performance based on a different template/variation.
2. COT was found to enhance performance and possibly consistency
3. Proper reporting and visualization is important to convey an accurate picture of the consistency of LLMs.
4. Regex and Logprobs prove unreliable.

🤔 What does the mismatch rate of first-token probability VS text output say about the source of LLMs reasoning?