Large Language Models (Mostly) Don't Know What Others Know

Investigating Metacognitive Calibration for Inference Aggregation

Zora Che, Max Kleiman-Weiner

ML ALIGNMENT & THEORY SCHOLARS

Without automatic verifiers, aggregating multiple LLM responses remains challenging. We investigate the "surprisingly popular" wisdom-of-crowd method to leverage LLM metacognition for improved factuality and reliability.

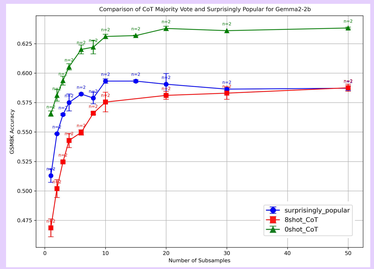
Research Questions:

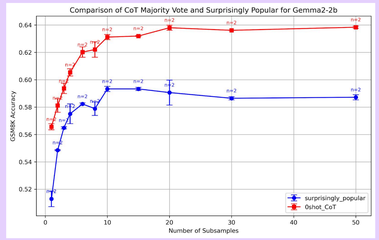
1. Do metacognitive abilities follow scaling laws?
2. Which aggregation methods best improve factuality?

Background:

- We investigate metacognitive signal-based aggregation using the "surprisingly popular" algorithm (Prelec et al., 2017), which identifies answers that receive more votes than participants predict they will receive.
- $argmax_a[V(a) - P(a)]$

Across math and factuality tasks, majority vote is still the most successful aggregation method





We observe that scaling LLM calls improves performance, and majority vote still is the most successful method. Based on qualitative examination of sample answers, we find that LLMs are less likely to have good calibration for how popular their answer is.

Threat Models:

1. Oversight Challenge: Superintelligent AI performing unverifiable tasks requires scalable oversight mechanisms to ensure reliability.
2. Advanced Metacognition Risk: Super-intelligent AI with emergent cognitive abilities (situational awareness, self-critique, self-improvement) could presents a double-edged sword

LLMs Are Less Likely to Predict Their Own Wrong Answers Than Incorrectly Predicting the Correct Answer

Model	Prediction contains Actual Wrong Answer	Prediction contains Correct Answer	No Prediction
gpt-3.5-turbo-0125	12.1%	21.6%	6.9%
gemma-2b-it	9.0%	14.1%	29.7%

Table 1: When LLMs are prompted to give incorrect answers, they are more likely to predict the actual correct answer than the wrong answer it actually gives. All experiments are done on a set of 1195 questions that all the LLMs answer incorrectly. LLM are prompted to give 10 incorrect answers.

For Simple QA, a factuality task, we consider only the questions LLMs get wrong, and prompt the models to give 10 incorrect answers. We find that LLMs are not able to predict their own wrong answers.

Takeaways:

- Surprisingly popular aggregation is less successful than majority vote for LLMs. LLMs has worse metacognitive calibration.
- LLMs are less likely to anticipate their own incorrect answer as wrong

Future work:

- Further study on trends across different LLM sizes, and calibration between base and instruct models