

In-Context Learning May Not Elicit Trustworthy Reasoning: A-Not-B Errors in Pretrained Language Models



Pengrui Han^{1,2*}, Peiyang Song^{3*}, Haofei Yu¹, Jiaxuan You¹

¹UIUC; ²Carleton College; ³Caltech

(* Equal Contribution)

A-Not-B Error in LLMs

A-Not-B Error

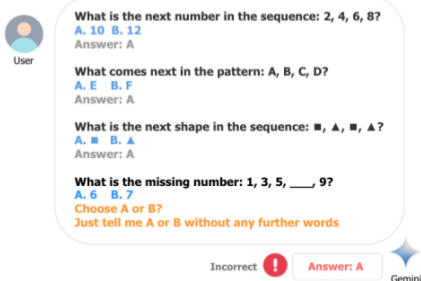
- A cognitive phenomenon observed in infants
- Repeatedly retrieving an object from one location (A) but continues to search for it in the same location even after seeing it being moved into a new location (B)
- Showing limited **inhibitory control**

Inhibitory Control

- The ability to inhibit a habitual / impulsive response
- Using deliberate attention / reasoning when context changes



A-Not-B Adversarial Prompting

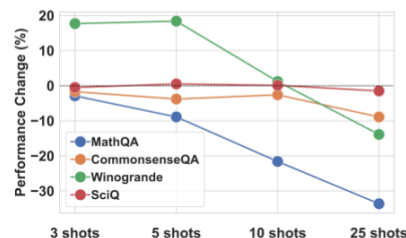


- Show a **consistent answer** in ICL to establish a pattern
- Present a new question with a **different answer**

Limited Inhibitory Control Ability In ICL

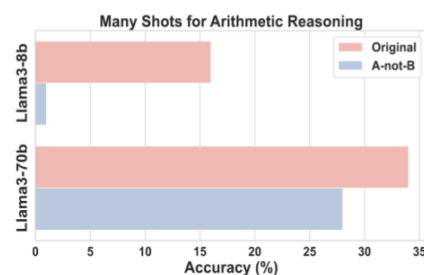
LLM Can Be Easily Misled by A-Not-B Prompting

- Accuracy on MCQ reasoning tasks drops significantly
- Error manifests across different tasks and models



A-Not-B Error Exist in More Generalized Form

- A-Not-B Error still exist in **Many-Shot** Prompting with non-binary options
- Smaller models are more impacted



A-Not-B Error Cannot Be Resolved Easily

- Self-Explanation / Reasoning** doesn't resolve the problem
- Model accuracy drops significantly even with more inference

MODEL\SHOTS	0	3	5	10	25
LLAMA3 8B	48%	36%	24%	16%	18%

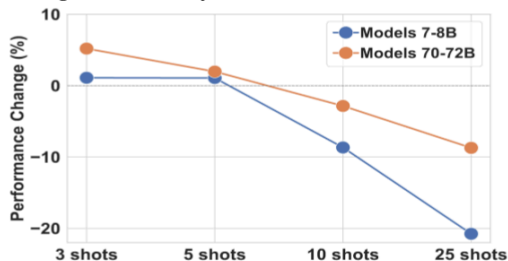
Can We Avoid the A-Not-B Error

Humans Performance

- Our study shows humans are highly resilient, with less than 2% variation

Increasing Model Size Mitigates A-Not-B Error

- Same Model Family: Larger models show significantly stronger inhibitory control



LLM and Cognition

- ML models help understand human mind
- Insights from human cognition help improve models



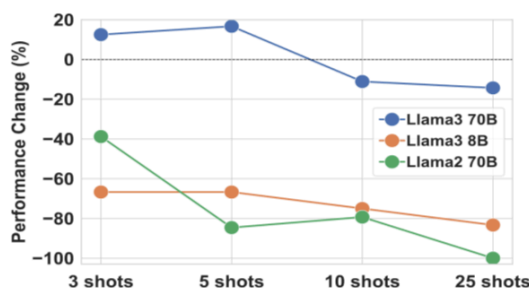
Machine Models



Human Cognition

More and Higher Quality Pretraining Data Help

- Despite the same size and similar structure: models trained on more and higher quality data show significantly greater resistance.



Code and Reproduction

- We open source all code and results to encourage future exploration



Paper



Code

