PhD Open Days

The Importance of Being Statistically Earnest:
A Critical Re-evaluation of GSM-Symbolic

PhD in Computer Science and Engineering

Dominika Agnieszka Długosz (✉ dominika.dlugosz@tecnico.ulisboa.pt)

EMNLP2026 BUDAPEST

CAN LLMS REASON?

Not according to Apple.

> We hypothesize that [...]
>
> **current LLMs cannot perform genuine logical reasoning;**
>
> they replicate reasoning steps from their training data.

— Imam Mirzadeh et al., 2025:
GSM-Symbolic: *Understanding the Limitations of Mathematical Reasoning in Large Language Models*

But how strong is their proof?

GSM-SYMBOLIC: DON'T LET LLMS CHEAT

1. Take a GSM8K question (Cobbe et al. 2021).
2. Swap names, numerical values, objects, currencies...
3. Repeat 50 times for 100 questions.

GSM8K question
Olivia has $23. She bought five bagels for $2 each. How much money does she have left?

GSM-Symbolic variants
Sarah has €30. She bought eight donuts for €3 each. ...
Ana has £24. She bought three lollies for £6 each. ...
Sophie has...

x50

x100

5000 questions. 1 new benchmark.

25 LLMs. 25 accuracy drops.

One thing missing.

You have guessed it —

DO YOUR STATISTICS!

Our choice: **Generalised Linear Mixed Model** (GLMM).

Model 1

$$Correct \sim Variant + (1 \mid Id)$$

— grouping by template, does the given LLM fail more on GSM-Symbolic?

Model 2

$$Correct \sim Variant + \gamma_c + (1 \mid Id)$$

— does it (still) fail more if we account for the number load*?

*GSM-Symbolic used systematically larger numbers than the original GSM8K.
Correct: single outcome · **Variant**: GSM8K vs GSM-Symbolic · γ_c : number load · **Id**: question template id

...because

ACCURACY ONLY TELLS HALF THE STORY

and it may not even be the relevant half

8 / 20
LLMs show a significant accuracy drop on GSM-Symbolic (Model 1)

4 / 20
still do after correcting for the number load (Model 2)

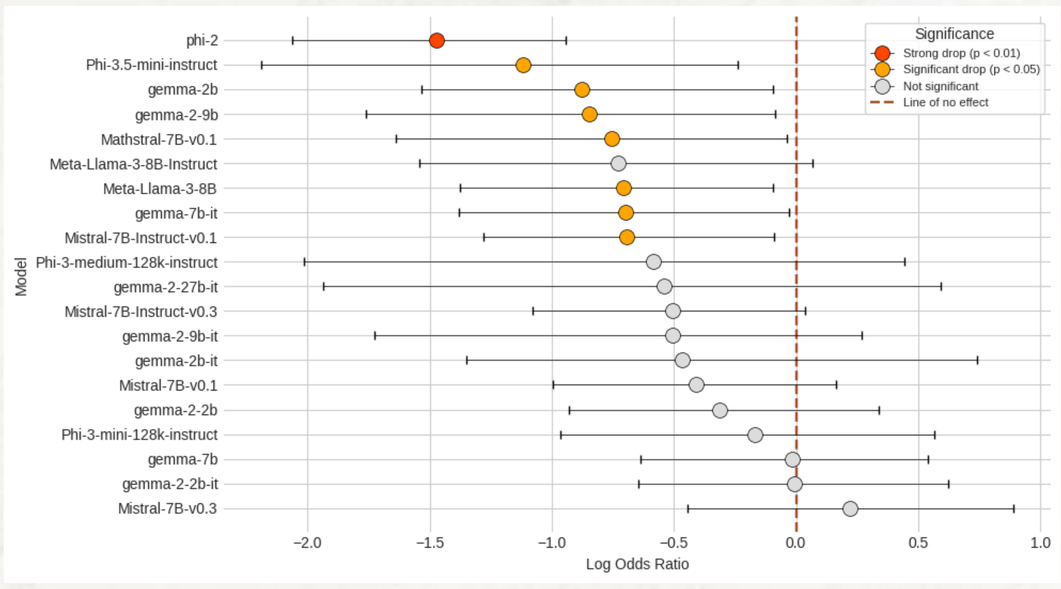


Figure 1: Model 1. Effect size of GSM-Symbolic (log-OR). 8 of 20 results significant.

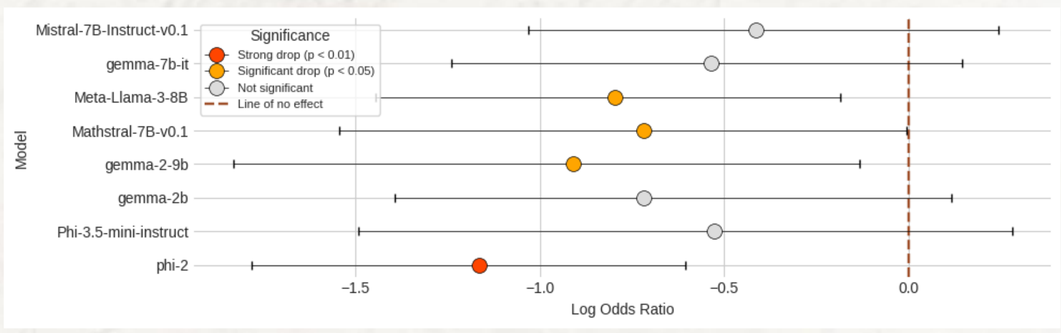


Figure 2: Model 2. Number-load-corrected effect size of GSM-Symbolic (log-OR) for the 8 LLMs with previously significant accuracy drops. 4 remain significant.

TÉCNICO UNIVERSIDADE DE LISBOA
UNIVERSIDAD DE GRANADA
ellis

Supervisors:
Arlindo L. Oliveira *(IST, INESC-ID)*
Natalia Díaz-Rodríguez *(U. de Granada)*

PhD in Computer Science and Engineering

phdopendays.tecnico.ulisboa.pt