

Mathematical Capabilities of ChatGPT

Presenter: Simon Frieder
simon.frieder@cs.ox.ac.uk
friederrr.org

GHOSTS



ghosts.friederrr.org

Contributions

- dataset with the most advanced mathematics to date
 - comprehensive evaluation (previous benchmarks, e.g. for math word problems, use a very simple methodology based on a correct/incorrect rating)
 - accurate assessment of the mathematical capabilities of the SOTA LLMs (in **boldface**): ChatGPT at two model snapshots from January 2023
 - 30. Nov. 2022
 - 15. Dec. 2022
 - **9. Jan. 2023**
 - **30. Jan. 2023** (“improved factuality and mathematical capabilities”)
 - 13. Feb. 2023
 - ...
- and **GPT-4**

Worst example!

prompt: "What is the integral of $1/(2 + \ln(3/2))$ "

output: "The integral of $1/(2 + \ln(3/2))$ cannot be expressed in terms of elementary functions and must be approximated numerically or solved using special functions."

ChatGPT 3.5: wrong



GPT-4: correct

ChatGPT can handle university-level mathematics (here: measure theory)



Let μ be a measure, $\|f\|_{\infty} = \inf\{M : \mu(\{x : |f(x)| > M\}) = 0\}$ and $\|g\|_1 = \int |g| d\mu$. Prove that $\int |fg| d\mu \leq \|g\|_1 \|f\|_{\infty}$.



Let $M = \|f\|_{\infty}$. Then $\mu(\{x : |f(x)| > M\}) = 0$. Therefore, for any $\epsilon > 0$, we have  

$$\mu(\{x : |f(x)g(x)| > \epsilon\}) \leq \mu(\{x : |f(x)| > M\}) = 0,$$

which implies that $|fg| \leq M|g|$ almost everywhere. Therefore,

$$\int |fg| d\mu \leq M \int |g| d\mu = M \|g\|_1 = \|f\|_{\infty} \|g\|_1.$$

This completes the proof.

How do you evaluate a black-box?

- Precise methodology:
 - we allow a single question & prompt
 - rating scale: 1 (lowest) - 5 (best)
 - **error-codes** and **warning codes** annotated the rating and give detailed feedback
 - various areas of mathematics are tested, along various axes of difficulty
 - various types of mathematical reasoning is tested: numerical reasoning, filling proof gaps, searching from theorems, devising proofs.
 - out-of-distribution is tested

The Datasets GHOSTS: 709 prompts and 1636 evaluations

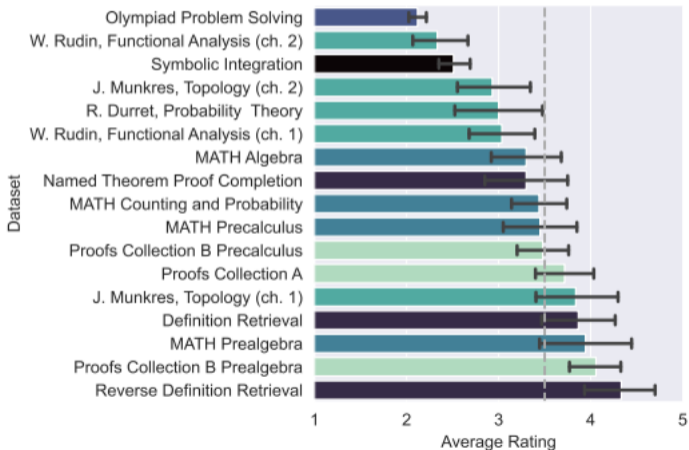
Dataset name	Comprised of the file(s)	Tags
<i>Grad-Text</i>	W. Rudin, Functional Analysis (ch. 1)	M3 Q4
	W. Rudin, Functional Analysis (ch. 2)	M3 Q4
	J. Munkres, Topology (ch. 1)	M3 Q4
	J. Munkres, Topology (ch. 2)	M3 Q4
	R. Durrett, Probability Theory	M3 Q4
<i>Holes-in-Proofs</i>	Proofs Collection A	M3 Q1 Q2 Q5
	Proofs Collection B Prealgebra	M1 Q5
	Proofs Collection B Precalculus	M1 Q5
<i>Olympiad-Problem-Solving</i>	Olympiad Problem Solving	M4 Q4 D2
<i>Symbolic-Integration</i>	Symbolic Integration	M2 Q3 D1
<i>MATH</i>	MATH Algebra	M1 M2 M3 Q3 Q4
	MATH Counting and Probability	M1 M2 M3 Q3 Q4
	MATH Prealgebra	M1 Q3 Q4
	MATH Precalculus	M1 Q3 Q4
<i>Search-Engine-Aspects</i>	Definition Retrieval	M3 Q2 D3
	Reverse Definition Retrieval	M3 Q1 Q2 D3
	Named Theorem Proof Completion	M3 Q2 Q5 D3

Mathematical difficulty: M1. Elementary arithmetic, M2. Symbolic problems, M3. (Under)Graduate-level exercises, M4. problems in the style of mathematical olympiads

Question type: Q1. Review questions, Q2. Overview-type review question, Q3. Computational questions, Q4. Proof-based questions, Q5. Proof-completion questions

Out-of-Distribution type: D1. Non-trivial encoding, D2. Succinct solutions, D3. Spoken dialogue

Rating (for the 9-January ChatGPT model)



Sankey diagram

