

Mathematical Capabilities of (Chat)GPT

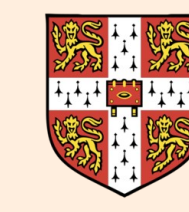


S. Frieder*, L. Pinchetti, A. Chevalier, R.-R. Griffiths, T. Salvatori,
T. Lukasiewicz, P.C. Petersen, J. Berner

*corresponding author (simon.frieder@cs.ox.ac.uk, <https://friederrr.org>)



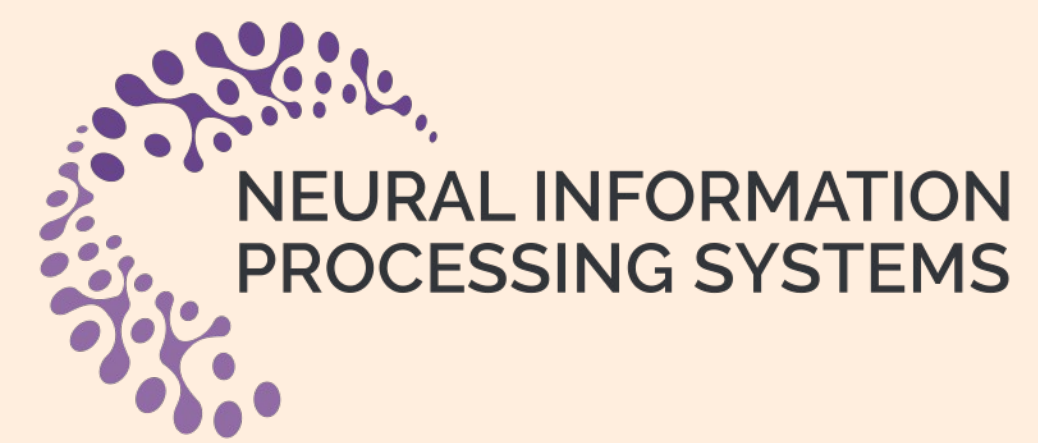
UNIVERSITY OF OXFORD
PRINCETON UNIVERSITY



UNIVERSITY OF CAMBRIDGE
Caltech



universität wien
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology
VERSES

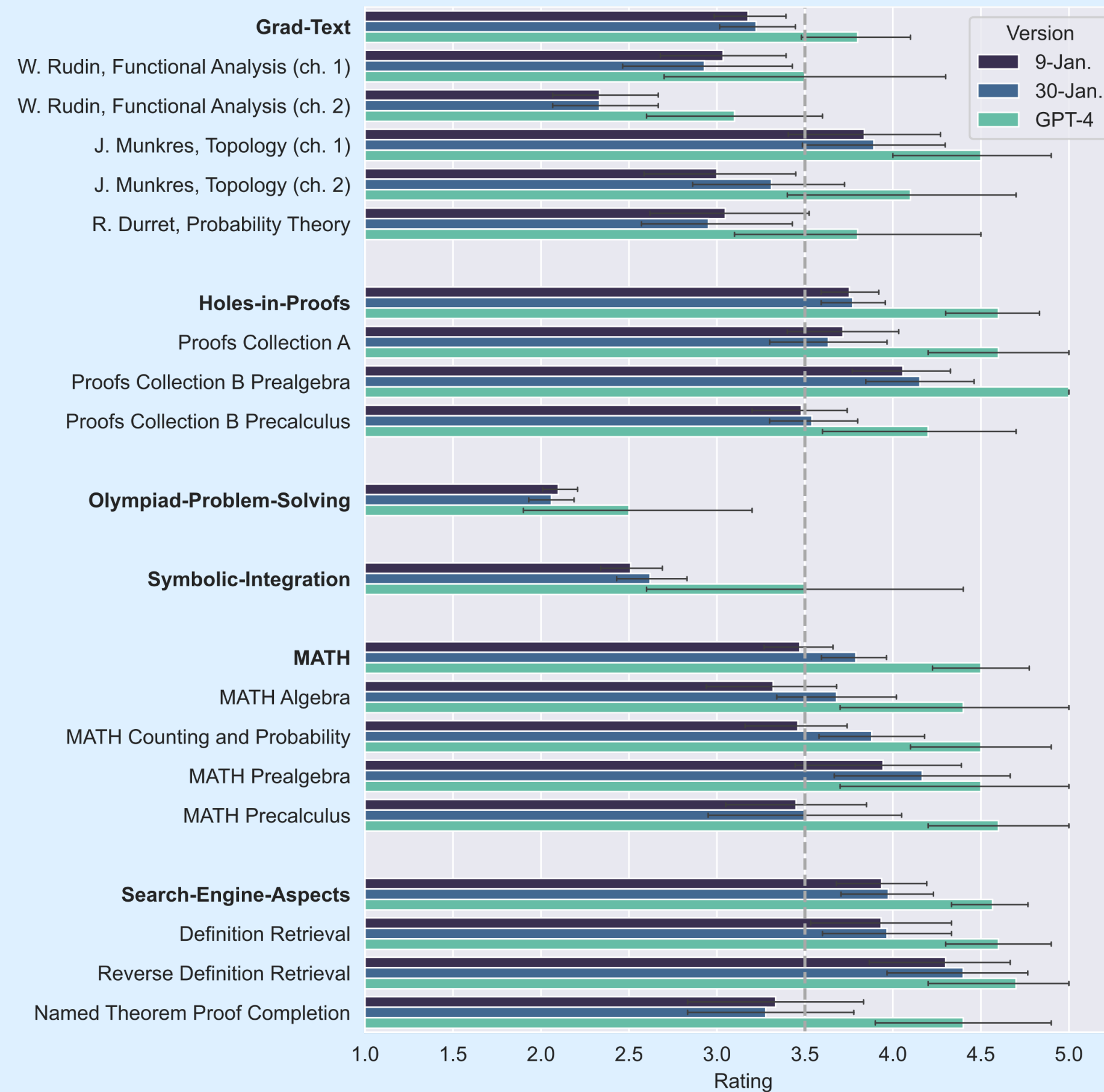


The GHOSTS > miniGHOSTS > microGHOSTS dataset

- **1600+ datapoints** on various question types (devise proofs at different levels of difficulty, fill in holes in existing proofs, finding mathematical definitions and named theorems, integration etc.)

- **three versions of (Chat)GPT:** 9-January-2023, 30-January-2023 (“improved factuality and mathematical capabilities” according to OpenAI’s release notes), and GPT-4

• GHOSTS:



Average ratings on the GHOSTS dataset for ChatGPT and on the miniGHOSTS dataset for GPT-4. We show the rating for each file from each of the six subdataset that make up (mini)GHOSTS, as well as the averages on the entire subdatasets. We consider 3.5 to be a passing grade. All models struggle on competitive mathematics.

- **Leaderboard: Use microGHOSTS and test your own model**

Bad Examples

ChatGPT (version 9-January-2023) gets it wrong:

"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."

GPT-4 also gets things wrong:

"prompt": " $x^2 + y^2 = x^2y^2$ has no integral solutions besides $x = y = 0$."
"output": "I'm afraid that's not entirely correct. The equation $x^2 + y^2 = x^2y^2$ actually has infinitely many integral solutions, including $x = y = 0$. To see this, note that the equation can be rewritten as $x^2(1 - y^2) + y^2 = 0$. This can be further rearranged to $x^2 = \frac{y^2}{y^2 - 1}$. This equation holds for all integer values of y such that $y = -1, 0, 1$, as these values result in $x = 0$ (which is an integer). However, for $|y| > 1$, the equation has no integral solutions for x , since the right-hand side will be a fraction. Therefore, the only integral solutions are $(x, y) = (0, -1), (0, 0), (0, 1)$."

What kind of (*grad-level, Olympiad*) mathematics can LLMs do and how do we measure it?

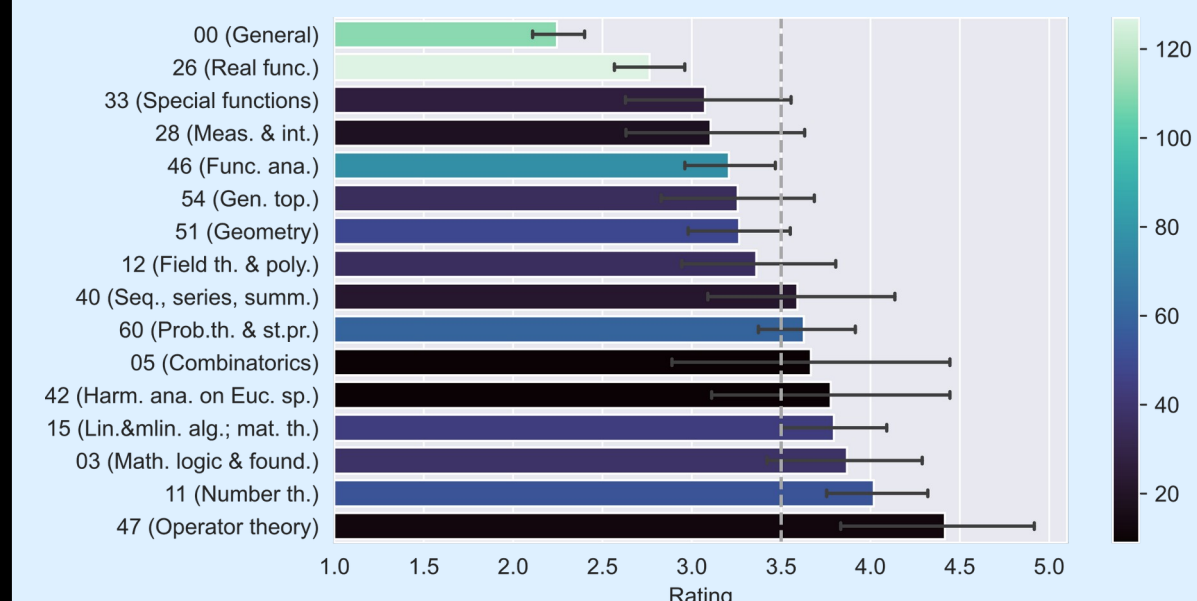
Novel Benchmark for Mathematics in Natural-Language

- traditional benchmarks are binary
- we introduce a fine-grained benchmark: our error codes, warning codes tells us not just **THAT** it failed but **HOW** it failed

Version Improvements, Results, and a Good Example

- Sankey diagram shows: grades are increasing
30-Jan-23 version (“improved factuality and mathematical capabilities”) doesn’t show a big increase

- use MSC codes to identify math subdomains

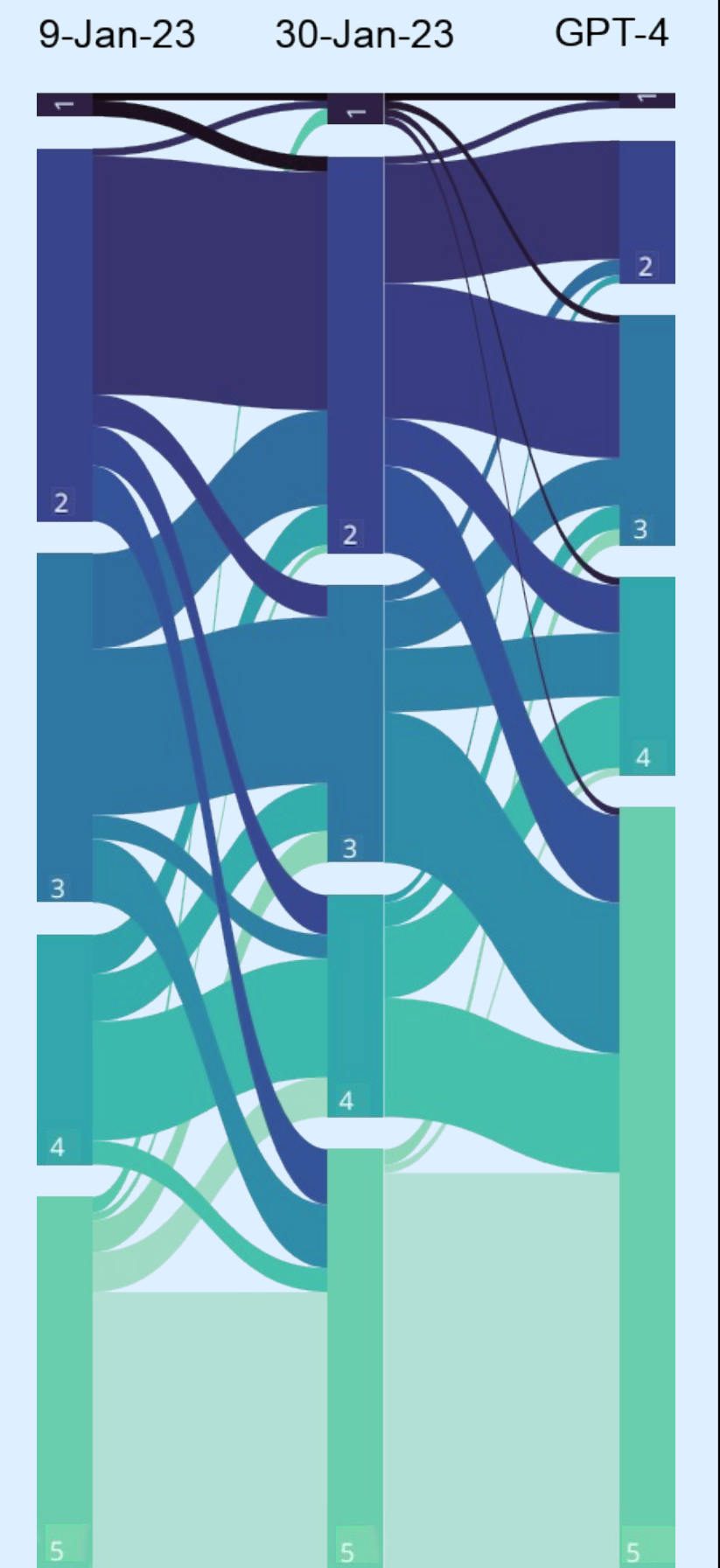


- GPT-4 outperformed on specialized tasks (integration) by specialized models (w.r.t. MATH dataset)

- LLMs have best performance when used as a **search engine**, see example below

GPT-4 gets it right:

"prompt": "How do you call a topological space, which has a countable dense subset?",
"output": "A topological space that has a countable dense subset is called a "separable space". The term "separable" comes from the idea that the points in the space can be "separated" or distinguished from each other using only countably many rational numbers. A classic example of a separable space is the set of real numbers, \mathbb{R} , with the standard topology, because the set of rational numbers, \mathbb{Q} , is countable and dense in \mathbb{R} . Another example is the space of continuous functions on the interval $[0,1]$ with the topology of uniform convergence, which is separable because the set of polynomials with rational coefficients is countable and dense in this space."



Check out the latest version at [arXiv.org](https://arxiv.org/abs/2301.13867)
<https://arxiv.org/abs/2301.13867>, or on our website
ghosts.friederrr.org (see QR code) for more information, diagrams about this project as well as media reports on our research

