HARDMath2: A Benchmark for Applied Mathematics Built by Students as Part of a Graduate Class

James V. Roggeveen, Erik Y. Wang, David Ettel, Will Flintoft, Peter Donets, Lucy S. Nathwani, Nickholas Gutierrez, Anton Marius Graf, Siddharth Dandavate, Arjun Nageswaran, Raglan Ward, Ava Williamson, Anne Mykland, Kacper K. Migacz, Yijun Wang, Egemen Bostan, Duy Thuc Nguyen, Zhe He, Marc L. Descoteaux, Felix Yeung, Shida Liu, Jorge García Ponce, Luke Zhu, Yuyang Chen, Ekaterina S. Ivshina, Miguel Fernandez, Minjae Kim, Kennan Gumbs, Matthew Scott Tan, Russell Yang, Mai Hoang, David Brown, Isabella A. Silveira, Lavon Sykes, Ahmed Roman, William Fredenberg, Yiming Chen, Lucas Martin, Yixing Tang, Kelly Werker Smith, Hongyu Liao, Logan G. Wilson, Alexander Dazhen Cai, Andrea Elizabeth Biju, Michael P. Brenner

School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

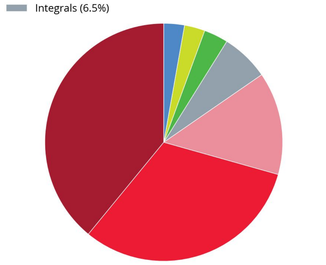
Why HARDMath2?

- SOTA math benchmarks focus on **exact solutions** or **formal proofs**
- Physics and engineering mathematical problems rely on **approximations**, **asymptotics**, and **perturbation methods** that are underrepresented in current evaluations

We provide:

- 211 **original** graduate-level applied math problems across 5+ domains, targeting approximation reasoning & asymptotic methods
- A custom evaluation framework that avoids LLM-as-a-judge
- Unique pedagogical approach combining benchmark creation & education

The HARDMath2 Dataset



Sample HARDMath2 Problems

Boundary Layers
Consider the differential equation:

$$\epsilon y'' - xy + x^3 y = 0$$

Find a uniformly valid approximation to the solution with boundary conditions $y(0) = A$ and $y(1) = B$ in the limit $\epsilon \ll 1$.

Nonlinear Partial Differential Equations
Consider the Korteweg-de Vries (KdV) equation:

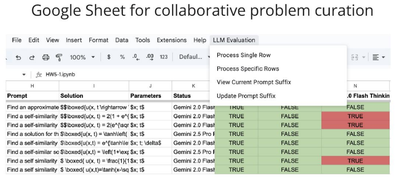
$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0$$

Find a soliton solution in the limit $t \Rightarrow \infty$.

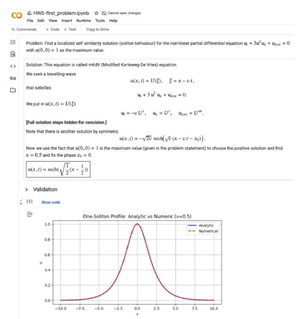
Pedagogical Approach

- Built by students as part of a Harvard graduate applied math class
- Homework involved creating problems that SOTA models **could not solve**
- Students' problems were evaluated by custom Google Sheet integration, giving **real-time feedback** on problem difficulty
- Peer verification for correctness of solutions
- By competing against models, students identified **critical failure modes**

Google Sheet for collaborative problem curation



Integration with automatic parser & eval



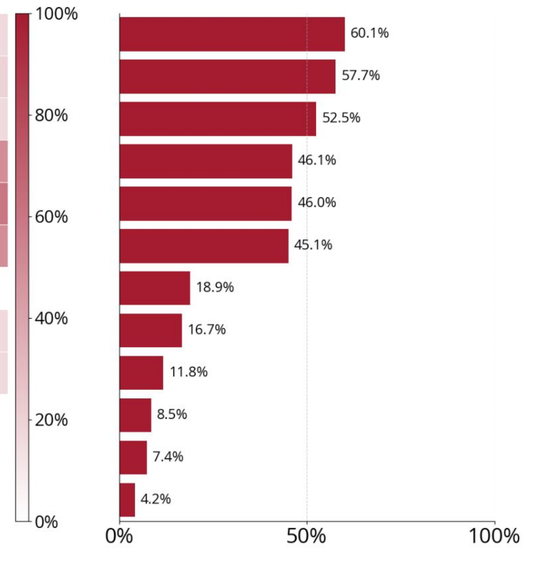
Solution and validation in Colab notebook

Model Performance

Performance Breakdown by Category

	Asymptotic Series	Boundary Layers	Integrals	Nonlinear ODE	Nonlinear PDE	WKB	Other
Gemini 2.5 Flash Thinking	20.0	77.8	71.4	40.0	45.0	61.5	16.7
Gemini 2.5 Pro Preview	25.0	72.6	78.6	14.3	44.3	60.0	20.0
o3	50.0	64.8	76.9	40.0	35.7	53.8	16.7
o4-mini	25.0	53.6	76.9	60.0	23.1	50.0	50.0
o3-mini	0.0	44.6	78.6	40.0	37.8	50.0	60.0
o1	0.0	47.3	71.4	50.0	31.5	55.6	50.0
DeepSeek V3	0.0	22.5	46.2	42.9	6.7	18.2	0.0
Llama 4 Maverick	0.0	18.3	76.9	28.6	3.3	11.5	16.7
Claude 3.7 Sonnet	33.3	4.2	61.5	28.6	1.7	24.0	16.7
Gemini 2.0 Flash	0.0	1.3	33.3	33.3	9.3	13.6	0.0
GPT-4.1	0.0	1.4	38.5	16.7	5.1	14.3	0.0
GPT-4o	0.0	0.0	23.1	28.6	2.3	6.2	0.0

Overall Solve Rate



Evaluation Scheme

- HARDMath2 implements a **custom LaTeX to SymPy parser** to avoid LLM-as-a-judge
- Numerical evaluation accommodates equivalent forms of expressions without needing symbolic simplification
- Model and reference solutions evaluated numerically and compared to determine equivalence up to some tolerance
- Problems judged based on equivalence to ground-truth solution **without** subjective partial credit scores

Conclusions

- **Overall Performance:** reasoning models struggle, hovering around 50-60% accuracy.
- Non-reasoning models perform significantly worse (<10% on many categories).
- **Specific Weaknesses:** Models are weakest on Nonlinear PDEs (requires physical ansatz selection) and Boundary Layers (requires matching inner/outer solutions).

Common Failure Modes

- Incorrect dominant balances
- Solving to a different asymptotic order than specified
- Not fully solving the problem by leaving off key matching steps to get final answer (see below sample)
- Failure to follow formatting guidelines

Sample Gemini 2.5 Pro Output for Boundary Layer

The leading order behavior of $y(x)$ as $x \to 0^+$ is given by $y(x) \sim Ax^{4/3}\exp(\frac{3}{2x^{1/3}})\cos(\frac{3\sqrt{3}}{2x^{1/3}} - \phi)$ where A is a positive constant, and ϕ is a phase constant. Since the constants A and ϕ are fixed by the boundary conditions and are non-zero, this gives the leading behavior.

Final Answer: The final answer is $\boxed{y(x) \sim Ax^{4/3}e^{3/(2x^{1/3})}\cos\left(\frac{3\sqrt{3}}{2x^{1/3}} - \phi\right)}$

Note: The LLM makes some attempt to narrow down what the constants are, but it does not solve for them completely. It mentions that the initial conditions fix the constants, but it does not list the equations it would use to solve for these constants, nor does it solve them.

Paper link:

Harvard John A. Paulson **School of Engineering** and Applied Sciences

NEURAL INFORMATION PROCESSING SYSTEMS