SorryDB: Can AI Provers Complete Real-World Lean Theorems?

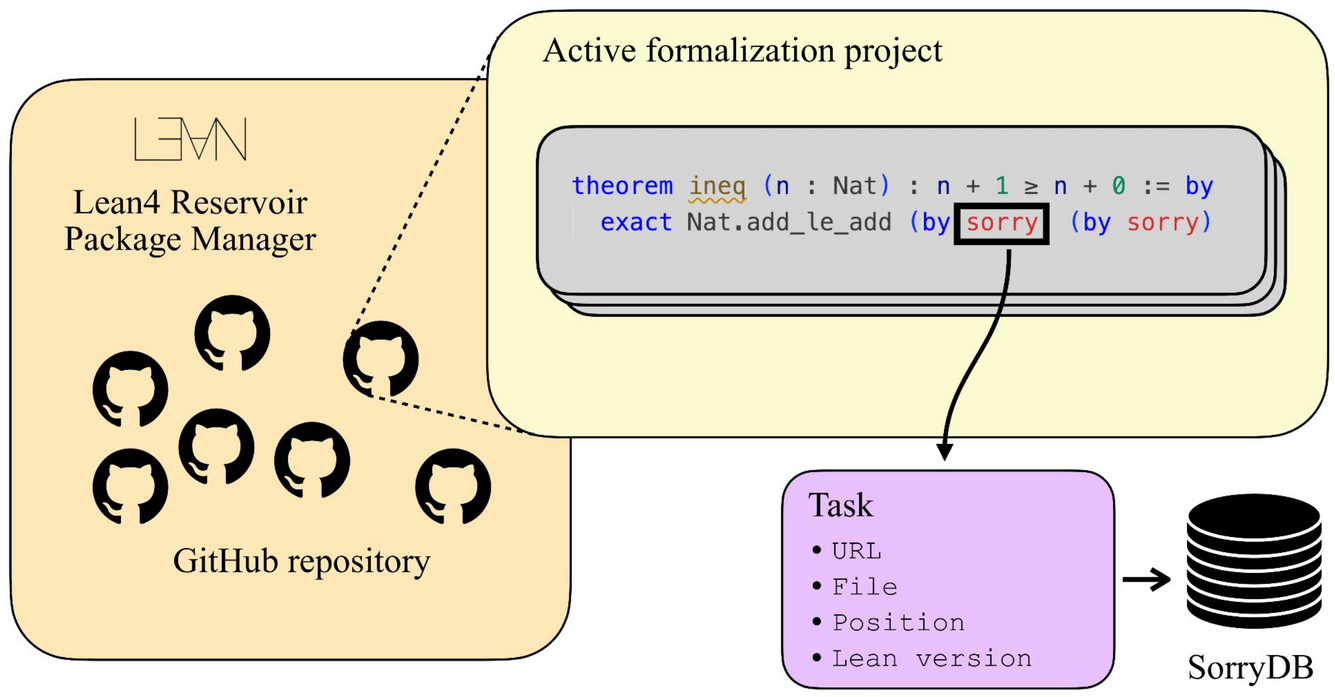
Austin Letson[1], Leopoldo Sarra[1], Auguste Poiroux[2,3], Oliver Dressler, Paul Lezeau[5,6], Dhyan Aranha[7,12], Frederick Pu[8], Aaron Hill, Miguel Corredera Hidalgo[9], Julian Berman[10], George Tsoukalas[11], Lenny Taelman[7]

[1]Axiomatic AI,[2]Math, Inc., [3]EPFL, [5]The London School of Geometry and Number Theory, [6]Imperial College, [7]University of Amsterdam, [8]University of Toronto,[9]ENSEIRB-MATMECA, INP-Bordeaux, [10]Columbia University, [11]The University of Texas at Austin, [12]Côte d'Azur University,

The SorryDB Dataset

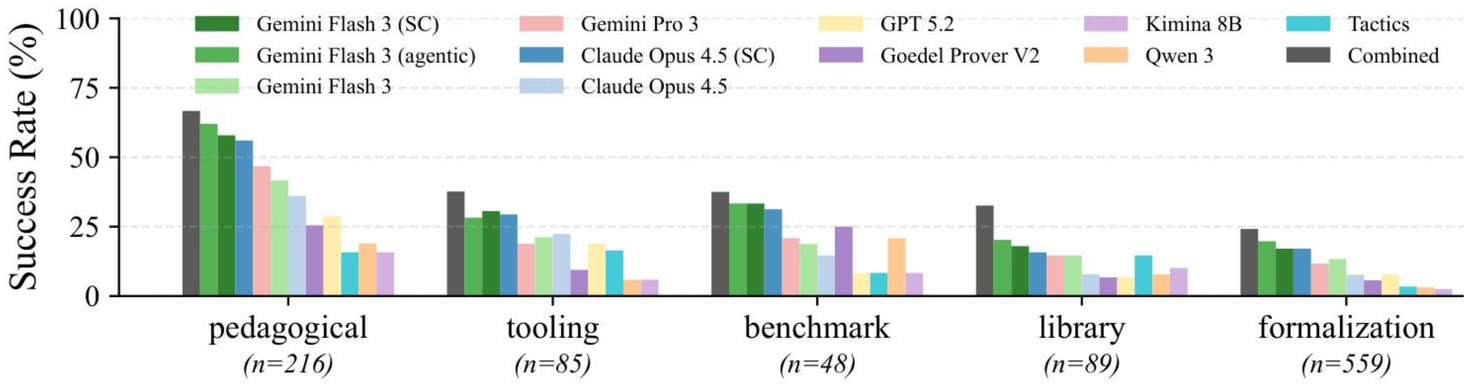
SorryDB is a dynamically-updating benchmark of open Lean tasks drawn from real world formalization projects on GitHub.

Hillclimbing the SorryDB benchmark will yield tools that are aligned to the community needs, more usable by mathematicians, and more capable of understanding complex dependencies.



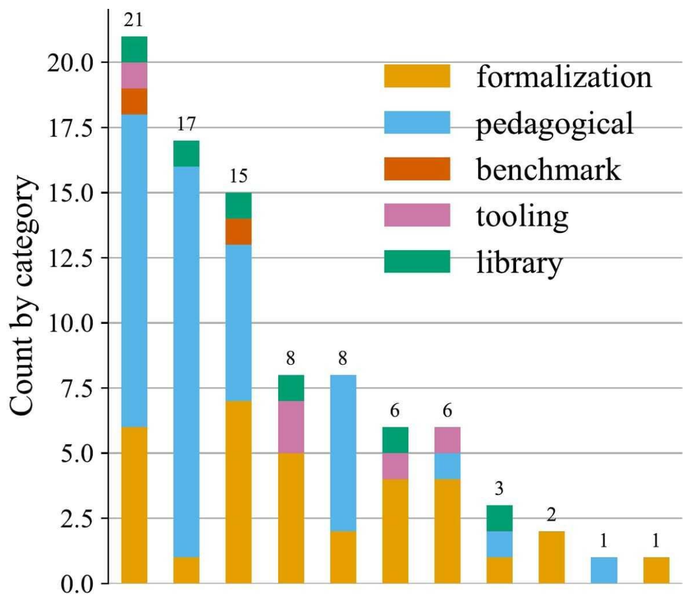
Evaluation

Approach	Pass@1	Pass@32
Deterministic		
Trivial	2.1 %	N/A
Tactics	8.4 %	N/A
General-purpose LLM		
GPT 5.2	6.2 %	13.2 %
Claude Opus 4.5	7.8 %	15.4 %
Gemini Flash 3	10.8 %	20.5 %
Gemini Pro 3	11.0 %	20.5 %
Qwen 3	5.0 %	8.1 %
Specialized LLM		
Kimina Prover 8B	1.0 %	6.6 %
Goedel Prover V2 32B	2.7%	11.3 %
Iterative		
Claude Opus 4.5 (SC)	27.1 %	N/A
Gemini Flash 3 (SC)	27.9 %	N/A
Gemini Flash 3 (agentic)	30.3 %	N/A
Combined	35.7 %	



Evaluation by repo category

Pedagogical and benchmark repos are easier than larger formalizations



Evaluation results

- Deterministic strategies already show baseline performance
- Iterative strategies outperform pass@k
- General purpose LLMs outperform specialized provers

Try it here!

Provers are complementary

The most powerful prover does not cover all solved tasks.

Upset plot of sorries solved by each model. Each column represents the number of theorems solved by all the provers marked in the table with a black circle, and not solved by the others. Deterministic tactics solve some sorries that none of the LLM-based approaches could solve.