BraveATA

Benchmarking Broad and Verifiable End-to-End Automated Theoretical Analysis of Large Language Models

Dingzirui Wang, Xuanliang Zhang, Keyan Xu, Qingfu Zhu, Wanxiang Che

NEURAL INFORMATION PROCESSING SYSTEMS

01 THE BENCHMARK

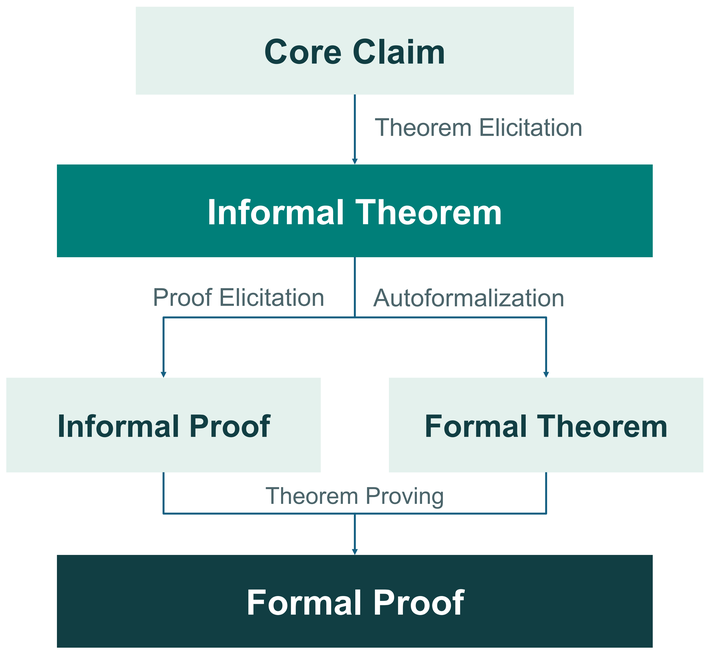
Benchmark for LLM Analysis

Can LLMs turn research claims into verifiable theory?

178 source papers | **11** directions | **4** tasks

One claim per paper, selected by five experts.
ICLR, ICML and NeurIPS 2025.

Four Tasks with Expert-Labeled Inputs



Each task uses expert-labeled inputs.
Formal tasks include the Lean 4 source header.

02 MODEL EVALUATION

Formal Proof is the Bottleneck

2.2% Best formal-proof Pass@16
4 of 178 theorems
Up to 16 attempts per theorem

14 models evaluated across six families

Main Performance on Benchmark

Model	Informal Theorem[1]	Formal Theorem[2]	Formal Proof[3]	Informal Proof[1]
Gemma4 31B	0.245	5.6	0.0	0.705
GPT-oss 117B-A5.1B	0.248	7.9	0.0	0.612
Qwen3.5 397B-A17B	0.252	8.7	0.0	0.696
DeepSeek-V3.2 671B-A37B	0.245	5.6	0.0	0.655
GPT-5.4	**0.287**	**9.0**	1.7	**0.790**
Claude Opus-4.6	0.273	6.7	**2.2**	0.729

[1] Rubric agreement with references (0–1).
[2] Lean-verified statement equivalence (%).
[3] At least one verified proof in 16 attempts (%).

Where Models Fail

85.7% models proved no theorem

Faithful mathematical abstraction is difficult.
Proofs often omit required assumptions.

03 END-TO-END STUDY

Automated Mechanism Explore

GPT-5.4 explores new claims with three agents.

Planner	Sets proof goals and assumptions.
Formalizer	Revises Lean statements using compiler feedback.
Extractor	Turns compiled statements into natural-language claims.

Claims Retained at Each Stage

173 Generated → **28** Expert-selected → **12** Lean-verified

6.9% of generated claims are Lean-verified

Main rejections: limited research value (76) and lack of novelty (42).

QR Code for More Detail

Paper | Code & Benchmark | Homepage