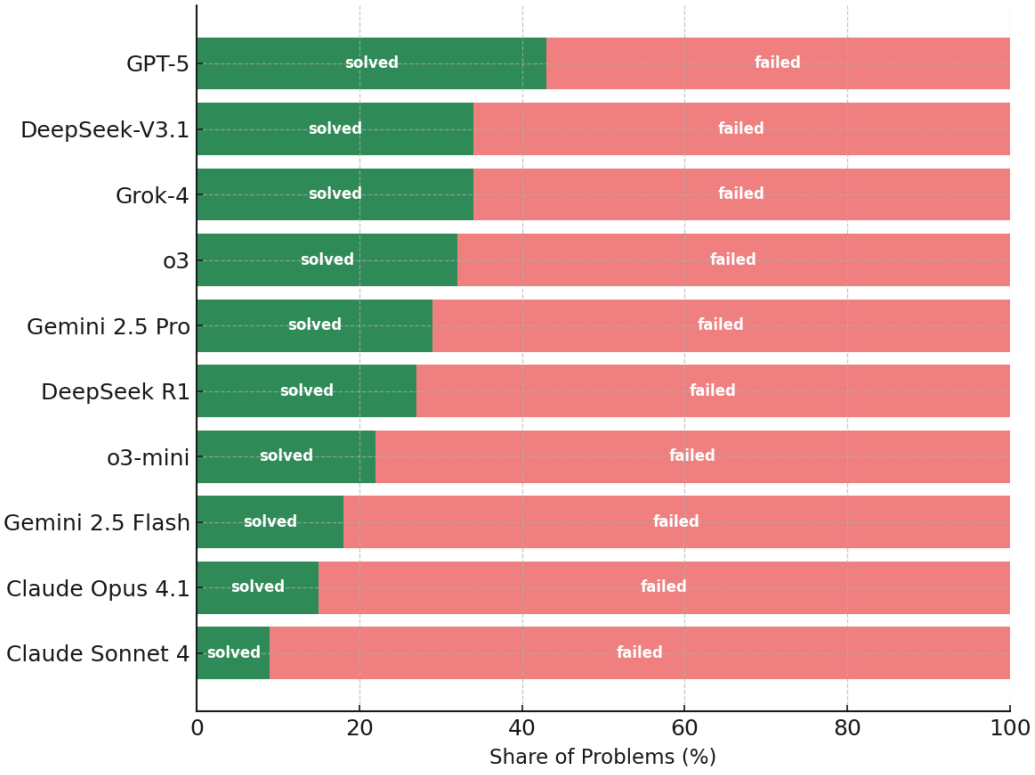
ScienceBench: A New AI Benchmark of Research-Level Math Problems

100 PhD-level research mathematics problems put GPT-5, Claude, Gemini, Grok, and others to the test.

Bochum, Germany – September 1, 2025 – math.science-bench.ai today released its first benchmark of large language models on research-level mathematics. While all models fail the majority of the problems posed by professional researchers, the results reveal wide variation in performance.

The benchmark evaluates 100 PhD-level research mathematics questions submitted by 37 contributors. Ten leading models were tested, including GPT-5 (OpenAI), Claude (Anthropic), Gemini (Google DeepMind), Grok (xAI), and DeepSeek. Among the active models, GPT-5 solved 43% of the problems, while most others remained below one third.

“These problems are PhD-level problems, requiring a deep theoretical understanding,” said Prof. Christian Stump, coordinator of the project. “The results show the limitations of current AI systems at this professional level.”



About math.science-bench.ai
The project evaluates the capabilities of AI systems in research-level mathematics. It is coordinated by Prof. Christian Stump from Ruhr-Universität Bochum in collaboration with international researchers, aiming to measure progress, identify limitations, and foster collaborations.

Contact
Prof. Dr. Christian Stump
Email: contact@science-bench.ai