

GPT-5.5: OpenAI's New Model Discovers a Ramsey Number Proof Verified in Lean

Kabui, Charles

2026-04-28

[Read at ToKnow.ai](#)

GPT-5.5: OpenAI's New Model Discovers a Ramsey Proof Verified in Lean

82.7% on Terminal-Bench 2.0, novel Ramsey number proof formally verified

82.7% Terminal-Bench 2.0 score (GPT-5.4: 75.1%)	35.4% FrontierMath Tier 4 (GPT-5.4: 27.1%)	\$5/M Input token price with 1M-token context window
--	---	---

April 28, 2026

ToKnow.ai

OpenAI released GPT-5.5 on April 23, 2026. It scores 82.7% on [Terminal-Bench 2.0](#) (complex command-line workflows), up from GPT-5.4's 75.1% and well ahead of Claude Opus 4.7 at 69.4%. On FrontierMath Tier 4, which tests graduate-level mathematical reasoning, it reaches 35.4%, up from GPT-5.4's 27.1%. The standout claim: an internal version of GPT-5.5 with a custom research harness discovered a new proof about off-diagonal [Ramsey numbers](#), a core

problem in combinatorics where new results are rare and technically difficult. The proof was formally verified in [Lean](#), a proof assistant that checks logical correctness mechanically. GPT-5.5 ships with a 1M-token context window at \$5 per million input tokens and \$30 per million output, matching GPT-5.4 latency on NVIDIA GB200 and GB300 NVL72 systems.

The Ramsey result is what separates this release. Models writing better code and scoring higher on benchmarks are expected increments. An LLM producing a novel mathematical argument that survives formal machine verification is qualitatively different. The Lean check matters: without it, the claim could be dismissed as hallucination or a restated known result. For developers, the pricing story is also notable. [Artificial Analysis](#) reports GPT-5.5 delivers frontier-level coding at roughly half the per-task cost of competing models, because it reaches answers in fewer tokens.

If AI can produce proofs that pass automated checkers, the bottleneck for mathematical discovery shifts from human intuition to the speed of proof search. Formal verification tools like Lean, not raw model scale, become the gating factor for trusting AI-generated mathematics.

Read More: OpenAI flags evidence of memorization on SWE-Bench Pro, a concern explored in [a recent audit of AI agent benchmarks](#).

Sources:

- [OpenAI: Introducing GPT-5.5](#)
- [OpenAI: GPT-5.5 Ramsey Number Proof \(PDF\)](#)
- [OpenAI Developer Docs: Using GPT-5.5](#)
- [Artificial Analysis: Intelligence Benchmarking](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. Read more: [/terms-of-service](#)