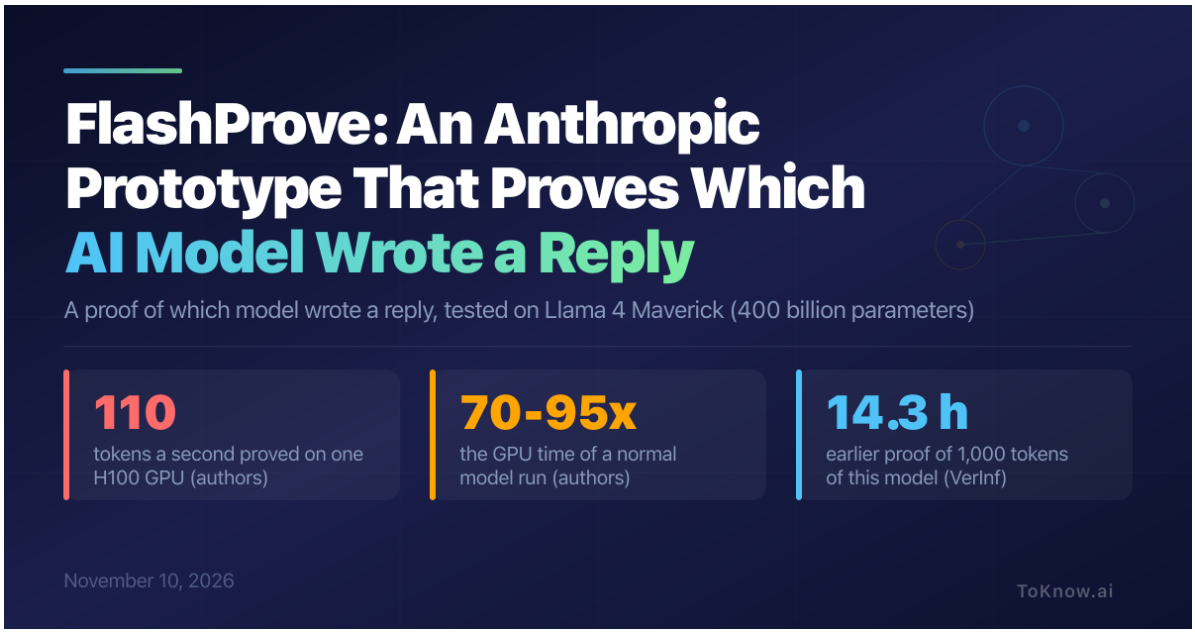
FlashProve: An Anthropic Prototype That Proves Which AI Model Wrote a Reply

Kabui, Charles

2026-11-10

Read at ToKnow.ai



FlashProve is an Anthropic research prototype that lets an AI company attach a mathematical proof to each reply, showing which model wrote it. It is for people who buy AI through resellers, and for auditors and regulators, because a reseller could quietly swap in a cheaper model. The proof works like a notarized receipt for a secret recipe: anyone can check the stamp without seeing the recipe. The company first publishes a short fingerprint of its model's weights, the

numbers that make up the model. After each reply, a GPU builds a proof file, and anyone with the fingerprint, the prompt and the reply can check it on an ordinary computer in about 20 seconds. On Meta's Llama 4 Maverick, a model with 400 billion of these numbers, the authors report proofs at 110 tokens (word pieces) a second on one H100 GPU.

Reading the reply is not enough to catch a swap. UC Berkeley researchers found that text classifiers scored about 50%, no better than a coin flip, at telling compressed versions of open models from the originals. An earlier proof system, VerInf, needed 14.3 hours to prove 1,000 tokens of the same model on a desktop NVIDIA DGX Spark. FlashProve is far faster but still uses 70 to 95 times the GPU time of a normal model run. The buyer must get the fingerprint from a trusted party, not the reseller, and each proof covers at most 1,024 tokens.

Read More: An LLM Cannot Prove Its Identity by Naming Itself, on why a model's own answer to "who are you?" is not evidence of which model replied

Sources:

- Progress on provable inference for large language models (Anthropic)
- FlashProve known limits (Anthropic, GitHub)
- Anthropic found a way to prove which AI model answered you (The New Stack)
- Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs (UC Berkeley, arXiv)
- VerInf: zero-knowledge proofs of large-language-model inference (GitHub)

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