Weight-Space Programming: One AI Reads and Edits the Numbers in Other AI Models

Kabui, Charles

2026-11-09

Read at ToKnow.ai



MIT researchers report that an AI model can read, fix and write the numbers inside other AI networks without any training. Those numbers, called weights, decide what a network does. In a November 2 blog post, OpenAI's GPT-6 Astra got the weights of small networks as plain text, with no tools. It named the game each of four networks played. For a simulated car stuck against a ramp, it changed 3.5% of the weights, and the car then succeeded in every test

run. From an all-zero network it wrote a Breakout player that scored 517.4, against 25.3 for one trained on 500 million game steps, though it trailed in Seaquest.

With code and up to 50 test runs, Astra edited Qwen3.5-0.8B, a small open model its makers say is prone to thinking loops, where it repeats itself until its answer is cut off. Astra removed the loop pattern from 0.016% of the weights, lifting the score on MATH-500 (500 maths problems) from 7.9% to 48.1%. Neel Nanda, a researcher in mechanistic interpretability (tracing what parts of a network do), wrote that Astra is "surprisingly capable of doing old school mechanistic interpretability research rather than anything spooky". On Gemma 3 270M, which rarely loops, no edit moved the score by more than about 1%.

The Hugging Face account akhaliq rebuilt the edit and scored 51.0% on MATH-500, though 31.6% of answers were still cut off. A separate team's three Astra searches on the same model gained 5%, 27.3% and 29.45%.

Read More: Antidoom, Liquid AI's training fix for reasoning models stuck repeating text and Qwenputer, a project where a compiler, not training, sets every weight of layers added to a Qwen model

Sources:

- Sparks of Weight-Space Programming (authors' blog post)
- Edited Qwen3.5-0.8B model card with the Hugging Face re-run (Hugging Face)
- VentureBeat: MIT researchers used GPT-6 Astra to edit Qwen's weights
- Neel Nanda's reaction (X)
- Exploring Weight-Space Programming: Capabilities, Generalization, and Limitations (arXiv)

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