

AI Models Stop Following References After Two Links, and a Tiny Add-On Fixes It

Kabui, Charles

2026-10-05

 Read at [ToKnow.ai](#)

AI Models Stop Following References After Two Links, and a Tiny Add-On Fixes It

One tweaked layer unlocks 160-link chains in frozen models

99%

accuracy on 24-line chains

0.01%

share of weights trained

160

links followed after the fix

October 5, 2026

ToKnow.ai

Give a pretrained language model a chain of references to follow inside its prompt, with each line pointing to the one before it, and most models lose the thread after two or three links and start guessing. A Georgia Tech team tested thirteen open models from Qwen3, Llama, OLMo-3, and Gemma-3 and found they reliably follow only 1.4 to 3.6 links, a median of 2.2. More depth does not help: OLMo-3's 32-layer and 64-layer versions both manage about 2.6.

The fix is tiny. A rank-8 LoRA, a small add-on that adjusts one layer while every original weight stays frozen, raises Qwen3-8B from 15.5% to 99% exact accuracy on 24-line chains, using 65,537 new parameters, under 0.01% of the model.

That suggests the ability is already inside the model, needing only a nudge to surface. The authors trace the effect to a relay: each line passes its chain identity down a short run of middle layers, letting the frozen attention heads read further up. On multi-hop questions built from invented facts, the same trick lifted exact-match accuracy from 41.2% to 97.4%, and handled six-hop questions it never trained on.

The lesson is about timing, not size. The chains are a synthetic probe rather than a general reasoning test, but early-layer add-ons also added 9% to 18% exact-match accuracy on MuSiQue, a real multi-hop benchmark. Models stop extending a chain once it passes their middle layers, and placement matters: moving the fix one layer later in Qwen3-8B cut reach from 20.5 lines to 5.2.

Read More: [Attention Residuals: A Drop-In Fix for How Every LLM Stacks Its Layers](#)

Sources:

- [Transformers Stop Thinking Too Early, and a Tiny LoRA Fixes It \(arXiv\)](#)
- [Interactive demo: watch the relay move across the layers](#)
- [Project code \(GitHub\)](#)
- [Hugging Face paper page](#)

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