Thinking to Recall: Why Reasoning Helps LLMs Remember Facts They Already Know

Kabui, Charles

2026-03-19

Read at ToKnow.ai

Google Research found that reasoning models perform better on simple factual questions that don't need step-by-step logic, and their new paper explains why. Using Gemini 2.5 Flash, Gemini 2.5 Pro, and Qwen3-32B with reasoning toggled on and off, they identified two mechanisms. First, a computational buffer effect: the model uses generated reasoning tokens to perform hidden-state computation regardless of what those tokens actually say. Replacing

real reasoning traces with repeated "Let me think." filler still boosted accuracy from 20.6% to 26.2% on SimpleQA. Second, factual priming: the model recalls related facts during reasoning that act as a semantic bridge to the correct answer, similar to how recalling the first nine kings of Nepal helps recall the tenth. Extracting just the facts from reasoning traces and feeding them back as context (with reasoning off) recovered most of the performance gains, confirming the facts themselves do the work.

The catch is serious. When the model hallucinates facts during reasoning, those wrong facts poison the final answer. On SimpleQA, 41.4% of reasoning traces with all-correct intermediate facts produced right answers, versus only 26.4% for traces containing hallucinated facts. The same pattern held on EntityQuestions: 71.1% correct with clean traces, 32.2% with hallucinated ones. Filtering for hallucination-free traces at inference time improved accuracy by up to 12.2%, pointing toward a practical fix.

This reframes what chain-of-thought actually does. For simple factual questions, the benefit isn't logical decomposition, it's giving the model more computation cycles and a chance to self-prime its own memory. It also means reasoning isn't just occasionally unreliable: there's a specific mechanism where bad intermediate steps actively make the final output worse.

Sources:

- Thinking to Recall: How Reasoning Unlocks Parametric Knowledge in LLMs (arXiv)
- HuggingFace Paper Page (#2 Paper of the Day, March 11)
- Gekhman et al. (Google Research) Author Page
- SimpleQA-Verified Benchmark
