The Deterministic Horizon

When Extended Reasoning Fails and Tool Delegation Becomes Necessary

Dongxin Guo [1] Jikun Wu [2,3] Siu Ming Yiu [1]

[1]The University of Hong Kong [2]Brain Investing Ltd [3]Stellaris AI Ltd

Forty-Third International Conference on Machine Learning (ICML 2026)

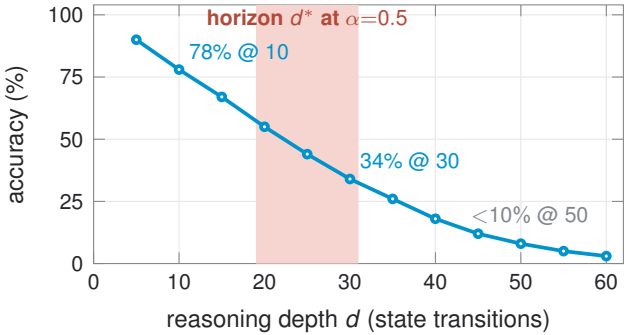
Longer reasoning makes *exact* tasks worse, not better

Past $\sim$20 exact reasoning steps, attention can no longer track state, so **delegate the computation to a tool**.

In **deterministic state tracking** ($\sigma_0 \to \tau$ via exact operators), *one wrong step is failure*, with no partial credit.

Puzzles BFS solves in $<$0.1 s make state-of-the-art models fail after *seconds* of deliberation.

The long reasoning **causes** the failure; it does not cure it.



Two explanations that make *opposite* predictions

PREFERENCE (alternative)

Models *could* track state but *prefer* short reasoning.

$\Rightarrow$ prompting / training should **recover** accuracy.

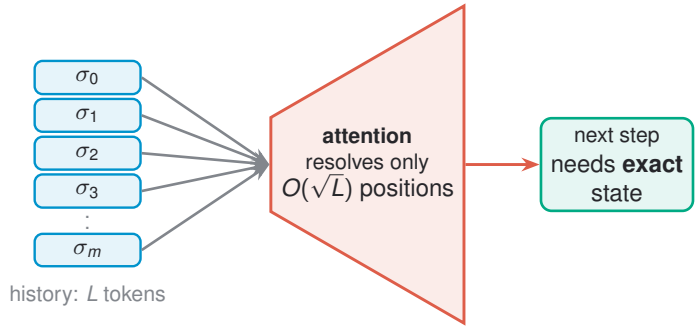
CAPABILITY (this work)

Attention *cannot* hold exact state past a limit: *decoherence*.

$\Rightarrow$ a hard **architectural ceiling** remains.

We lay out **two** contrasting predictions (fine-tuning and length prompts) and one **direct state-overlap measure**. The rest of the talk resolves them.

Why: attention squeezes all history through a narrow channel



Attention Bottleneck Theorem (matching construction)

$$|\mathcal{S}_{\text{track}}| \leq c \cdot 2^{H \log_2(L)\sqrt{d_h}} \Rightarrow O\big(H \log_2(L)\sqrt{d_h}\big) \text{ per layer}$$

Capacity is huge and is *not* the binding constraint; the **per-step** bits for exact tracking are what cannot fit. Bound is robust ($<20\%$ shift under $\pm 20\%$ perturbation).

So error grows with depth ⇒ a closed-form horizon

Attention entropy rises with position, so per-step error is *not* constant:

$$\epsilon(d) = \epsilon_0 + \gamma \frac{d}{L_{\text{eff}}}.$$

Survival probability then decays with a **quadratic depth term**:

$$P(\text{correct at } m) \leq \exp\Big(- m\epsilon_0 - \tfrac{\gamma m(m+1)}{2L_{\text{eff}}} \Big).$$

Fine-tuning ceiling (empirical): for $d > d^*$ the gain decays as $O(d^*/d)$; no training distribution escapes it.

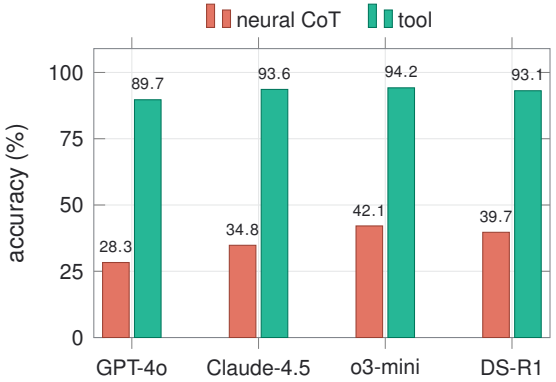
DETERMINISTIC HORIZON

$$d^* = \frac{-\epsilon_0 L_{\text{eff}} + \sqrt{\epsilon_0^2 L_{\text{eff}}^2 + 2\gamma L_{\text{eff}} \ln(1/\alpha)}}{\gamma}$$

primary suite $\boxed{d^* \in [19, 31]}$ at $\alpha = 0.5$

empirical fit $d^* \approx 0.314\sqrt{d_{\text{model}}}$
(7–8B: ∼19–20 | 70–72B: ∼28)

Evidence 1: tool delegation 76–94% vs neural 17–42%



12 models

5 conditions $\times$ 8 tasks $\times$ 3 runs

incl. SWE-Bench, WebArena, SQL-Multi

Cohen's $d \gg 1$ (very large); consistent across all model families.

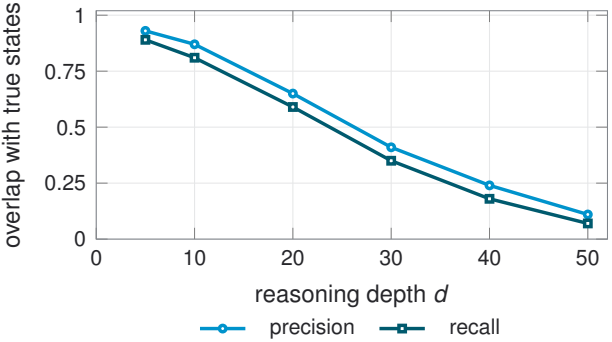
$\approx$**4**$\times$ faster per instance.

Evidence 2: it is architectural, and both predictions confirm

Discriminating test	Preference	Capability (ours)	Observed
Fine-tuning recovery	≥ 5 pp	<3 pp	**0.9 pp** ✓
"Reason longer" prompt	>10 pp	<2 pp	**0.8 pp** ✓

Preference levers (prompting, RLHF, fine-tuning)
cannot lift an architectural ceiling.

State overlap: the model drifts into states that don't exist

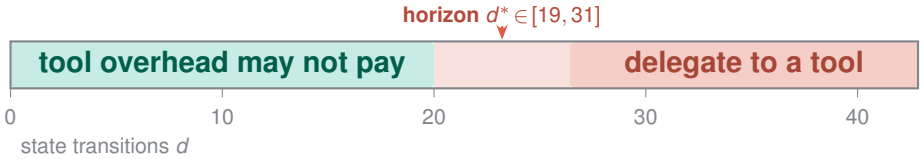


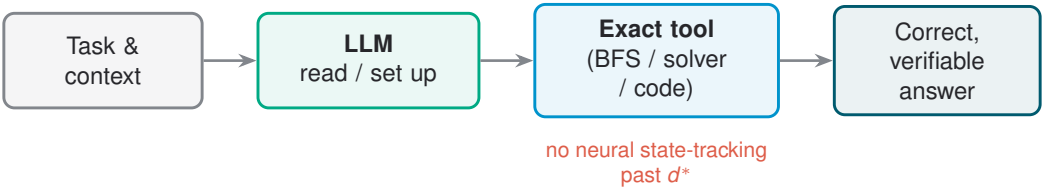
State-Space Jaccard compares the states the model claims with those its own operator sequence reaches.

Both precision *and* recall collapse together ($R^2 = 0.99$ fit), consistent with **capability** failure.

Precision exceeds recall at every depth, so we read SSJ as a measure of drift, not as a test between the accounts.

What to do: cross the horizon ⇒ delegate

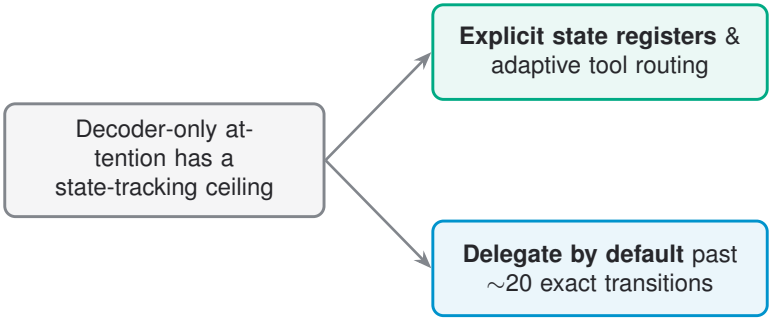




Delegate: program-state tracing, long puzzles, multi-hop tracking, many-table SQL joins.

Keep neural: short or approximate tasks (arithmetic, retrieval, summarization).

Takeaway



What would change the picture: an architecture that holds exact state past the horizon, or tool-free reasoning that matches **76–94%** at depth $> d^*$. Until then, near **∼20 transitions** the reliable choice is to delegate.

Thank you.