One Bit at a Time

Universal Per-Step Error Rates in LLM Compositional Reasoning

Anonymous Authors

Motivation

The Problem

When an LLM fails at multi-step reasoning, we lack a quantitative framework:

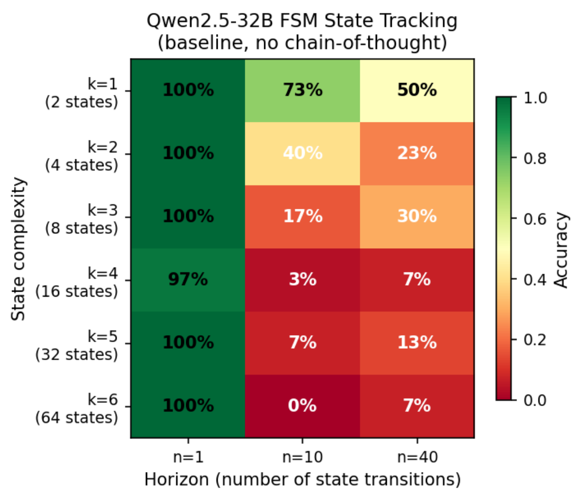
- Was the task too complex for a single step?
- Did errors accumulate across steps?
- Would a better model help, or more CoT tokens?

Current evaluations: monolithic benchmark scores
What we need: a predictive formula with interpretable parameters

The Cliff

Baseline (no CoT) on Qwen2.5-32B:

- $n=1$: near-perfect
- $n \geq 2$, $k \geq 3$: crashes to random
- A hard working memory wall



The Error Law

One Formula, One Parameter

Setup: Model carries k bits of state through n sequential steps.

Assumption: Each step independently corrupts each bit with probability $\epsilon = a \cdot k$.

$$\boxed{\eta = k \cdot \exp(-akn)}$$

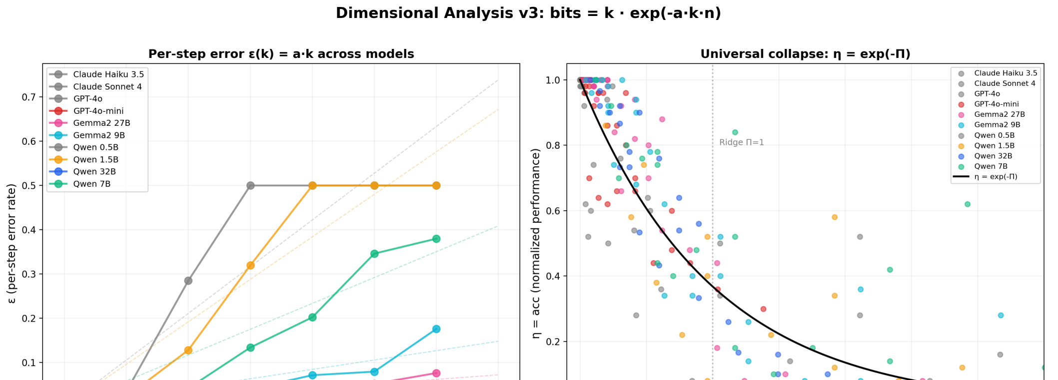
- η Expected bits correctly resolved (performance metric)
- k State complexity (bits per step)
- n Trace length (number of steps)
- a Per-step-per-bit error rate — one number per model

Universal Collapse

Define the dimensionless product $\Pi = akn$. Then:

$$\frac{\eta}{k} = \exp(-\Pi)$$

All models, all k, all n collapse onto a single curve.



Capacity Bound

Accuracy is non-negligible only when $akn \lesssim 1$:

$$k \cdot n < \frac{1}{a} \quad \Longleftrightarrow \quad \text{“reasoning horizon”}$$

Model	a	Horizon $1/a$
Claude Sonnet 4	0.00010	10,000
GPT-4o	0.00204	490
Qwen2.5-32B	0.00762	131
Qwen2.5-7B	0.06574	15

Reducing a by $2\times$ doubles the reasoning horizon as surely as doubling the context window.

Task Design

Random FSM Tracing

A controlled task isolating **pure state tracking**:

1. Random DFA with 2^k states, 4 input symbols $\{A, B, C, D\}$
2. Transition table: 4×2^k entries, uniformly random
3. Random start state + input sequence of length n
4. Model outputs the final state

Anti-memorization: Fresh random automaton per instance.
Parameters: $k \in \{1, \ldots, 6\}$, $n \in \{1, \ldots, 40\}$
CoT prompt: "Step i: $S_x + Y \to S_z$"

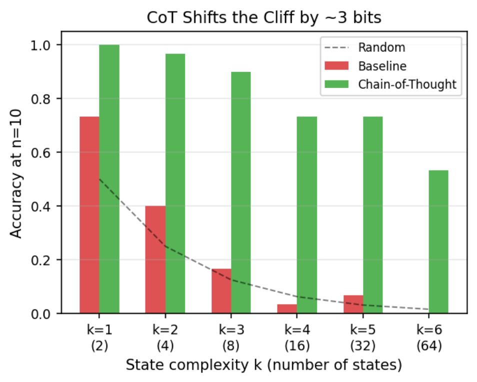
Results

Model Signatures: The a-Ranking

Model	a	R^2	Capacity $1/a$
Claude Sonnet 4	0.00010	0.999	10,000
Claude Haiku 3.5	0.00014	0.999	7,143
GPT-4o	0.00204	0.886	490
GPT-4o-mini	0.00705	0.958	142
Qwen2.5-32B	0.00762	0.817	131
LLaMA-3.1-70B	0.00810	0.891	123
Gemma2-27B	0.01042	0.903	96
Gemma2-9B	0.02108	0.860	47
Qwen2.5-7B	0.06574	0.724	15
Qwen2.5-1.5B	0.17294	neg.	—

a spans 650× across well-fit models. Ranking matches practitioner intuition exactly.

CoT vs. Baseline: Two Error Regimes



CoT: smooth exponential decay, $R^2 > 0.7$

Baseline: cliff at $k \geq 3$, $R^2 < 0$

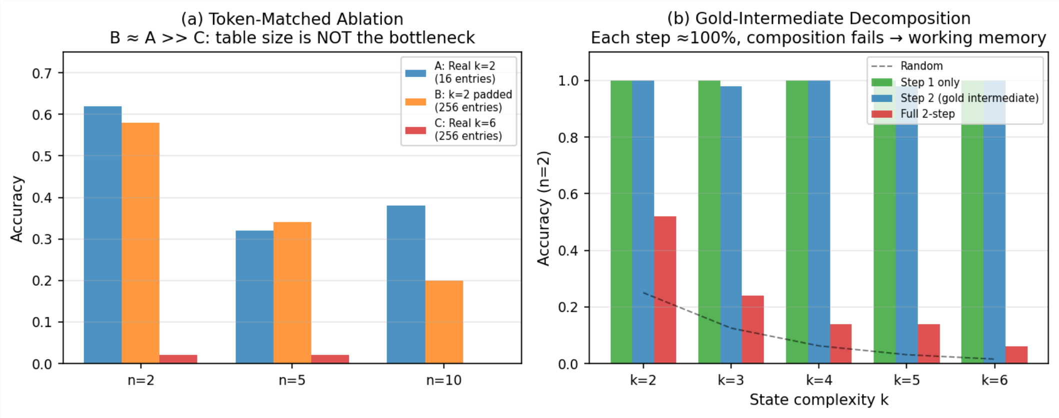
Key insight: CoT is not “smarter reasoning.”

It restructures errors from **correlated** to **independent**

— a **memory architecture** change.

Control Experiments

Two Controls: It's Working Memory, Not Retrieval



(a) Same token count, vastly different accuracy $\Rightarrow$ table size is not the bottleneck.

(b) Each step 100%, composed 6% at $k{=}6$ $\Rightarrow$ working memory is the bottleneck.

Cross-Task Generalization

Three Tasks, One Spectrum

Does the formula generalize beyond FSM tracing?

Task	Bit structure	CoT behavior	R^2
XOR chain	Independent	100% everywhere	N/A ($a \to 0$)
FSM tracing	Unstructured entanglement	Smooth $\exp(-akn)$	> 0.7
Mod arithmetic	Structured entanglement	Phase transition	-3.3

Key insight: The formula measures errors in **unstructured** state tracking.

- No entanglement (XOR): $a \to 0$ — formula trivially holds
- Opaque entanglement (FSM): model must track every bit — formula fits
- Structured entanglement (mod arith): model exploits algebra — formula breaks

Takeaways

Summary

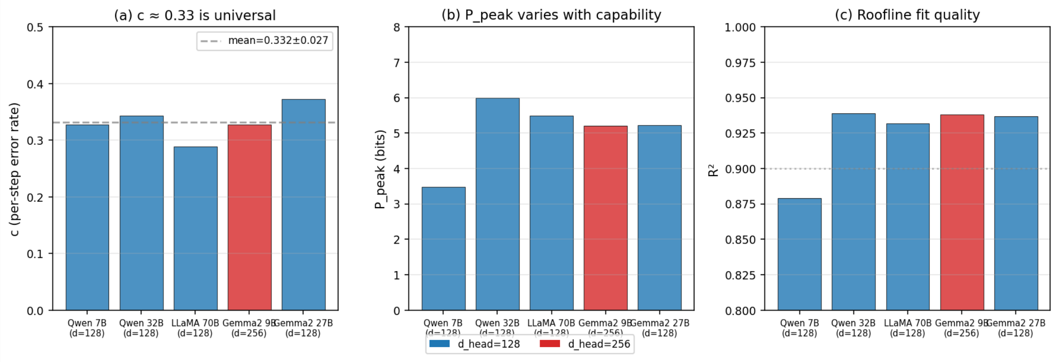
1. **One formula**: $\eta = k \cdot \exp(-akn)$ — fits 9/10 models, $R^2 > 0.7$
2. **One parameter**: a serves as a *model signature* ranking models across families
3. **Capacity bound**: $k \cdot n < 1/a$ defines the reasoning horizon
4. **CoT = memory architecture**: converts correlated $\rightarrow$ independent errors
5. **Bottleneck**: working memory, not retrieval (confirmed by two controls)
6. **Scope**: formula applies to unstructured state tracking (cross-task validated)

The per-step-per-bit error rate is the clock speed of compositional reasoning. Every model has one. Now we can measure it.

Thank you

Backup Slides

Parameter Universality Across Architectures



(a) Per-step error rate $c \approx 0.33$ is universal across 5 models.

(b) P_{peak} varies with model capability.

(c) Roofline fit quality: all $R^2 > 0.87$.