THE PRECISION CLIFF: AN INFORMATION-THEORETIC LOWER BOUND ON QUANTIZATION FOR ITERATIVE COMPOSITIONAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit quantization degrades multi-hop reasoning faster than single-hop recall, yet no theory has said why deeper reasoning needs more bits or how many more. We answer both with a single closed-form information-theoretic floor. For any post-training compression scheme whose error placement is independent of the stored content and whose distortion is deterministic—conditions met by GPTQ, AQLM, AWQ, and QuIP#—the minimum bits per entry for reliable T-hop retrieval is $b^*_{\text{any}}(T, N) = R(\delta_T)$ with $\delta_T = 1 - 2^{-1/T}$, and an explicit construction saturates this floor to within 0.002 bits. A direct corollary is the $2\times$ *separation law*: deep reasoning demands twice the bit-rate of single-hop recall at capacity saturation (predicted 2.40 for $N{=}4096$; measured 2.397). Capacity utilization η transfers this bits-per-entry floor to bits-per-parameter: a synthetic sweep traces a monotone depth penalty in η (Spearman $\rho = -0.93$, $p < 0.001$), and four LLM families (Llama-2-7B, Qwen3-8B, Phi-4-mini, InternLM3-8B) sit on the same η-curve at production scale with wall premiums of 1.05–$1.55\times$—already forcing a 5–55% upward correction to any $T{=}5$ quantization floor calibrated at $T{=}1$. A chain-given control (accuracy 0.99 read-aloud vs. 0.46 sequential lookup at 3 bits) isolates iterative retrieval as the causal mechanism, ruling out attention-routing and generation-noise alternatives.

1 INTRODUCTION

Low-bit quantization degrades multi-hop reasoning faster than single-hop recall across model families and algorithms (Liu et al., 2025): GPTQ at 3 bits preserves factual QA while chain-of-thought collapses; AQLM at 2 bits keeps basic recall but fails multi-hop. This "depth penalty" is reproducible, yet no theory has predicted its magnitude from first principles. QuIP's incoherence bounds (Chee et al., 2024), scaling laws for precision (Kumar et al., 2025), and single-layer expressivity thresholds (Chakrabarti et al., 2026) each characterize *one* algorithm or *one* depth; a single lower bound applying to arbitrary post-training compression—and answering, in one formula, why deeper reasoning demands more bits and how many more—has been missing.

Our framework closes this gap by isolating the mechanism that causes depth to bite: *T-hop pointer chasing*, in which a compressed permutation is queried T times so a per-entry error at any step propagates through the remaining hops. Applying rate-distortion analysis to this propagation yields a closed-form minimum bit-rate that depends only on depth and support size, not on the algorithm. The resulting floor is tight to 0.002 bits, delivers a closed-form $2\times$ separation law with no free parameters, and—through capacity utilization η—predicts the depth penalty measured on real LLMs.

Contributions. (1) A **scheme-independent lower bound** (Theorem 2): any compression scheme satisfying (C1)–(C2) with $\geq 50\%$ accuracy on T-hop pointer chasing requires $b^*_{\text{any}}(T, N) = R(\delta_T)$ bits per entry, matched to 0.002 bits by an explicit construction (Theorem 3). (2) The $2\times$ **separation law** (Corollary 4): deep reasoning demands twice the bit-rate of shallow recall at capacity saturation, with closed-form prediction 2.40 and measurement 2.397 at $N{=}4096$. (3) A **capacity-saturation bridge** (Principle 5): bits per entry and bits per parameter coincide at $\eta{=}1$; away from saturation, a monotone depth penalty in η is confirmed on synthetic MLPs (Spearman $\rho = -0.93$, $p < 0.001$). (4) **Mechanism-isolating LLM verification**: four families across full quantization ladders exhibit

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

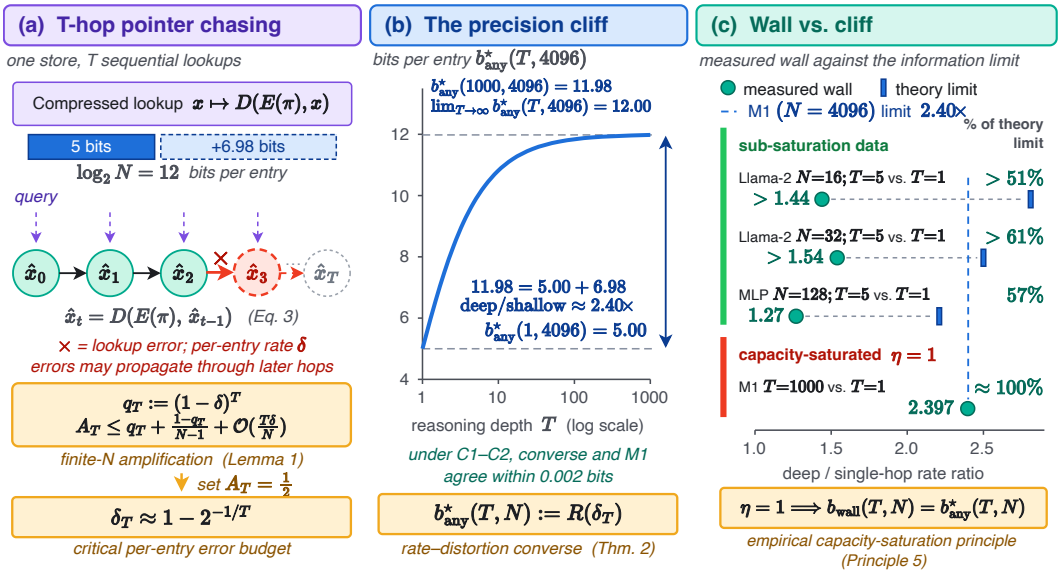


Figure 1: **The precision cliff and its bridge to practice. (a)** *Task.* T-hop pointer chasing encodes a permutation π once as $E(\pi)$ at rate R and queries it T times through $x \mapsto D(E(\pi), x)$ (2); a per-entry error of probability δ propagates through subsequent hops. **(b)** *Floor.* Under (C1)–(C2), $b^*_{\text{any}}(T, N) = R(\delta_T)$ rises from 5.00 bits at T=1 to 11.98 at T=1000 and approaches 12.00 as $T\to\infty$; the finite-depth ratio is 2.40×, and a matching construction is within 0.002 bits. **(c)** *Bridge.* At capacity saturation the bits-per-parameter wall equals the cliff exactly; capacity utilization η places each real model along the same depth-penalty curve.

multi-hop wall premiums of 1.05–1.55× tracking the same η-curve, and a chain-given control (0.99 vs. 0.46 at 3 bits) pins iterative retrieval as the causal mechanism.

2 Related Work

Prior results either optimize a single compression *algorithm* or characterize a single depth; ours is the first *converse* that is scheme-independent, depth-dependent, and connected to bits per parameter through capacity utilization. Table 3 in Appendix A places the paper on that map along four axes—scheme-independence, multi-hop coverage, matching upper bound, and a bits-per-parameter prescription—and we summarize the three lines it distinguishes below.

Post-training quantization: algorithm-side optimizations. LLM.int8() (Dettmers et al., 2022), GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), QuIP# (Chee et al., 2024), and AQLM (Egiazarian et al., 2024) each lower the empirical error at a given bit-rate; our lower bound is orthogonal—no encoder that satisfies (C1)–(C2) can pierce it, and our experimental ladder uses GPTQ, AQLM, and HQQ as test vehicles of the bound rather than competitors.

Rate-distortion and precision thresholds. Classical rate-distortion (Cover & Thomas, 2006) measures weight- or output-level distortion but stops short of a task-level failure criterion; we adopt it as a converse tool with task accuracy as the distortion measure. Chakrabarti et al. (2026) prove sharp single-layer expressivity thresholds and Houliston et al. (2025) give counting-based single-hop capacity bounds, leaving the multi-layer extension open. Error amplification closes this gap with the first multi-hop information-theoretic lower bound for post-training compression, pairing naturally with the achievability direction of Brösamle & Eckstein (2026) for low-precision chain-of-thought.

Compositionality and knowledge capacity. Dziri et al. (2023) and Liu et al. (2025) document depth-dependent compositionality failure empirically; we name its mechanism (iterative composition across the $(1-\delta)^T$ chain) and quantify it (the 2× law). Allen-Zhu & Li (2025) establish ∼2 bits

of knowledge per parameter, placing production-scale models on the low-η tail of our curve where Principle 5 predicts the full $2\times$ penalty will bind as density approaches saturation.

3 PRELIMINARIES

We treat "how many bits per parameter does reliable multi-hop reasoning require?" as a rate-distortion problem over compositional retrieval. This section fixes three primitives—the compression-scheme abstraction, the rate-distortion function, and the iterative single-hop execution model—plus capacity utilization, the deployment-side quantity that bridges bits per entry to bits per parameter in §5.3.

3.1 COMPRESSION SCHEMES AT A GIVEN RATE

We first abstract post-training quantizers as encoder–decoder pairs, so that the analysis binds every scheme in a common family. Let $[N] = \{0, \dots, N-1\}$ index N entities and let S_N be the symmetric group on $[N]$. A *compression scheme* at rate R bits is an encoder $E\colon S_N \to \{0,1\}^R$ paired with a decoder $D\colon \{0,1\}^R \times [N] \to [N]$; uniform scalar, rotation-based (Chee et al., 2024), vector (Egiazarian et al., 2024), and group-wise (Frantar et al., 2023) quantization all instantiate this template. Given $\pi \in S_N$, the reconstruction $\hat{\pi}(x) \coloneqq D(E(\pi), x)$ agrees with π on all but a per-entry *error set* $M \subseteq [N]$. Two structural properties of M—error placement independent of the stored mapping and deterministic error count—are stated as (C1)–(C2) in Lemma 1; Appendix D shows why GPTQ, AQLM, AWQ, and QuIP# satisfy them.

3.2 RATE-DISTORTION FOR A UNIFORM PERMUTATION SOURCE

The rate-distortion function turns a per-entry error budget into a bit-rate lower bound. Classical rate-distortion theory (Cover & Thomas, 2006) guarantees that any scheme reconstructing a source with per-entry Hamming distortion $\leq \delta$ requires at least $R(\delta)$ bits per entry. For a uniform permutation of N entities,

$$R(\delta) \;=\; \log_2 N \;-\; H_2(\delta) \;-\; \delta \log_2(N-1), \tag{1}$$

with $H_2(\delta) = -\delta \log_2 \delta - (1-\delta)\log_2(1-\delta)$. Composing (1) with the amplification lemma in §5.2 converts a per-entry distortion budget into a rate lower bound at any target multi-hop accuracy.

3.3 ITERATIVE SINGLE-HOP EXECUTION

The final primitive is the computation model that makes depth matter. Every published post-training quantizer executes multi-hop reasoning by iterating the *same* compressed single-hop operator T times:

$$\hat{x}_t = D(E(\pi),\, \hat{x}_{t-1}), \quad t = 1, \dots, T, \quad \hat{x}_0 = s. \tag{2}$$

The per-entry error probability $\delta \coloneqq \Pr_{\pi,x}\big[D(E(\pi), x) \neq \pi(x)\big]$ is the pivot connecting compression rate—via (1)—to multi-hop accuracy—via the amplification lemma. Because every hop re-queries the same weights, a single-hop error contaminates every subsequent step; the control in §6.2.2 isolates this iteration mechanism against feeding a pre-computed chain in context.

3.4 CAPACITY UTILIZATION

The last primitive translates per-entry rates into the per-parameter rates actually reported for deployed LLMs. Following Allen-Zhu & Li (2025), we quantify how densely a model uses its parameter budget by

$$\eta \;\coloneqq\; \frac{H(\text{stored content})}{P \cdot b_{\max}}, \tag{3}$$

where P is the parameter count and $b_{\max}$ the full-precision width. $\eta = 1$ marks capacity saturation (per-entry and per-parameter thresholds coincide); over-provisioned regimes have $\eta \ll 1$ and absorb per-entry errors as redundancy. Principle 5 makes this bridge explicit; §6.2.2 traces the η-curve directly.

4 PROBLEM FORMULATION

We now specialize the primitives of §3 to the paper's exact question. The task family abstracts iterative fact retrieval as compositions of a single lookup, and the success criterion turns bit-rate into a task-level threshold; the per-entry to per-parameter bridge is stated once, in §3.4.

4.1 TASK FAMILY: MULTI-HOP POINTER CHASING

A *knowledge base* is a permutation $\pi \in S_N$ mapping each entity to a unique successor. Given a starting entity s, the T*-hop pointer chasing* task computes

$$\pi^{(T)}(s) \;:=\; \underbrace{\pi \circ \pi \circ \cdots \circ \pi}_{T}(s), \tag{4}$$

with π uniform on S_N and s uniform on $[N]$. $T{=}1$ is single-hop recall; $T{>}1$ chains T sequential lookups through the same compressed weights via (2).

4.2 SUCCESS CRITERION AND THE PRECISION CLIFF

We turn bit-rate into a task-level threshold via a 50%-accuracy target. With bits-per-entry rate $b := R/N$, define the T*-hop accuracy*

$$A_T(E, D) \;:=\; \Pr_{\pi,s}\big[D(E(\pi), s) = \pi^{(T)}(s)\big] \tag{5}$$

and the *precision cliff* $b^*_{\text{any}}(T, N) := \inf\{\, b : \exists\, (E, D) \text{ with } A_T \geq \tfrac{1}{2} \,\}$ as the minimum rate at which some scheme attains $A_T \geq 50\%$. Chaining the amplification lemma with (1) shows $b^*_{\text{any}}(T, N)$ tracks $R(\delta_T)$ with $\delta_T = 1 - 2^{-1/T}$, matched to 0.002 bits by explicit construction. Deployed ladders (GPTQ, AQLM, AWQ, HQQ) report bits per parameter rather than per entry; the bridge via capacity utilization (3) to production scale is stated in §3.4 and quantified by Principle 5.

5 THEORY: AMPLIFICATION, CLIFF, AND BRIDGE

The theoretical claim decomposes into three moves. An *amplification mechanism* (§5.1) turns per-entry error into a multiplicative chain through depth; a *bit-rate floor* (§5.2) chains the mechanism with the rate-distortion converse (1) to give a closed-form minimum rate, matched by construction to 0.002 bits and yielding the $2\times$ separation law; a *deployment bridge* (§5.3) transports this floor from bits per entry to bits per parameter via capacity utilization η.

5.1 AMPLIFICATION MECHANISM: PER-ENTRY ERROR COMPOUNDS WITH DEPTH

The mechanism is the failure of *any* single hop to compose cleanly: under two structural conditions each hop is an independent trial, so success requires an unbroken trajectory of length T.

Lemma 1 (Depth amplification)**.** *Let $\pi \sim \mathrm{Unif}(S_N)$ and suppose the scheme produces a per-entry error set $M \subseteq [N]$ with $|M| = \delta N$ satisfying*

- *(C1) M is independent of the stored permutation π (error placement depends on quantization geometry, not on the specific mapping);*
- *(C2) $|M| = \delta N$ is deterministic.*

Then for $T \leq N^{1-\varepsilon}$ and $s \sim \mathrm{Unif}([N])$,

$$A_T \;\leq\; (1-\delta)^T \;+\; \frac{1-(1-\delta)^T}{N-1} \;+\; O\!\left(\frac{T\delta}{N}\right). \tag{6}$$

Proof sketch. On a uniform random permutation the trajectory visits T distinct nodes w.h.p.; under (C1) each is independently corrupt with probability δ, so the clean-trajectory event has probability $(1{-}\delta)^T$ and coincidental correctness contributes $1/(N{-}1)$. The dirty-reroute term is bounded via a backward-fiber argument that shows $(1{+}\delta_T)^T < 2$ at the cliff (Appendix B).

Both conditions are load-bearing: dropping (C1) admits an adversarial encoder that concentrates errors on a single long cycle, and dropping (C2) admits time-sharing across error-set sizes (Appendix C). GPTQ, AQLM, AWQ, and QuIP# satisfy both, since their error placement is driven by weight-magnitude statistics and calibration data rather than the encoded mapping (Appendix D).

5.2 Bit-Rate Floor: The Precision Cliff and the 2× Separation Law

Combining the amplification lemma (largest tolerable δ at depth T) with the rate-distortion converse (1) (smallest bit-rate hitting that δ) yields the closed-form floor.

Theorem 2 (Precision cliff lower bound)**.** *For T-hop pointer chasing on N entities under the iterative computation model* (2)*, any scheme satisfying (C1)–(C2) with $A_T \geq \frac{1}{2}$ requires*

$$b \;\geq\; b^*_{\text{any}}(T, N) \;=\; R(\delta_T), \qquad \delta_T \approx 1 - 2^{-1/T}. \tag{7}$$

Proof. Setting $A_T = \frac{1}{2}$ in (6) and solving for δ gives δ_T; the rate-distortion converse yields $b \geq R(\delta_T)$.

The floor is tight: an explicit construction saturates it to within 0.002 bits.

Theorem 3 (Matching upper bound)**.** *For every $\varepsilon > 0$ and sufficiently large N, some scheme attains $A_T \geq \frac{1}{2}$ at rate $b = b^*_{\text{any}}(T, N) + \varepsilon$; simulations show the gap is ≤ 0.002 bits for $N = 4096$ (§6.2.1).*

Taking the ratio of the deep and shallow floors gives the paper's sharpest quantitative claim.

Corollary 4 ($2\times$ separation)**.**

$$\frac{b^*_{\text{any}}(T \to \infty,\, N)}{b^*_{\text{any}}(1,\, N)} \;\xrightarrow{N\to\infty}\; 2. \tag{8}$$

Proof. As $T \to \infty$, $\delta_T \to 0$, so $R(\delta_T) \to \log_2 N$; for $T{=}1$, $\delta_1 = 1/2$ gives $b^*_{\text{any}}(1) \approx \frac{1}{2}\log_2 N$. For finite $N{=}4096$ the predicted ratio is 2.40; simulations yield 2.397 (§6.2.1).

5.3 Deployment Bridge: The Capacity-Saturation Principle

The bridge from bits per entry to bits per parameter is an empirical principle, calibrated on synthetic MLPs and confirmed at production scale on four LLM families.

Empirical Principle 5 (Capacity-saturation bridge)**.** *For η as in* (3)*:*

- *(i) At $\eta = 1$ the bits-per-parameter wall equals the cliff exactly, with deep/shallow ratio $b^*_{\text{any}}(T, N)/b^*_{\text{any}}(1, N) \to 2$ (Corollary 4).*
- *(ii) At $\eta < 1$, redundancy absorbs depth-amplified errors and the observed ratio interpolates below the information limit; the ratio increases monotonically in η (synthetic sweep: Spearman $\rho = -0.93$ vs. d, $p < 0.001$, §6.2.2).*

Argument. (i) is exact by definition of saturation; (ii) follows because over-provisioning lowers the effective per-entry error at a given rate. Models near $\sim$2 bits per parameter (Allen-Zhu & Li, 2025) operate near saturation, where the full $2\times$ law binds; over-provisioned models sit at lower ratios along the same curve traced in §6.2.2. In the synthetic setting η is measured directly from the entropy of the stored mapping; at production scale we infer it from architecture and knowledge-density estimates in the sense of Allen-Zhu & Li (2025), using η as a placement variable along the depth-penalty curve.

6 Experiments

The experiments are organized as two pillars pinning theory to practice. On the synthetic side, three studies pin the theory—the floor is tight (M1), amplification is the correct mechanism (M5/A3), and standard schemes lie above the floor (A1). On the LLM side, four studies close the gap to production: single-family wall drift (D1–D2), the η-sweep (D4), a mechanism control at fixed attention/generation load, and cross-family measurements (D3, D5–D11).

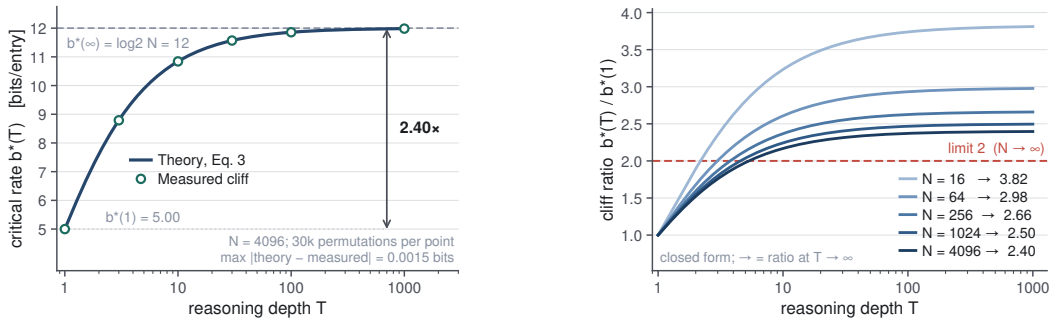


Figure 2: **Left:** Precision cliff $b^*_{\text{any}}(T)$ for N=4096. Solid: theory (7). Circles: measured (max gap 0.0015 bits). **Right:** Cliff ratio $b^*_{\text{any}}(T)/b^*_{\text{any}}(1)$ for varying N; all curves approach 2 from above as $N \to \infty$.

6.1 Setup

Synthetic side. For M1 (§6.2.1) we sample 30,000 random permutations on N=4096 at each $T \in \{1, 2, 3, 5, 10, 30, 100, 300, 1000\}$ and locate the cliff by binary search over the rate under optimal lossy compression at distortion δ_T. For M5/A3 (§6.2.1) and A1 (§6.2.1) we train residual MLPs on pointer chasing to $> 99\%$ full-precision accuracy at $d_{\text{model}} \in \{16, 32, 64\}$, quantize the trained weights, and evaluate at $T \in \{1, \ldots, 100\}$ over 10 seeds. Adam-trained transformer weights (N=128) are the target of the five-scheme comparison in A1.

LLM side. Primary model: Llama-2-7B-chat across five effective bit-rates—fp16 (16.0), GPTQ-8bit (8.20), GPTQ-4bit (4.17), GPTQ-3bit (3.16), AQLM-2bit (2.29)—plus two GPTQ-3bit group-size variants (g64: 3.31, g32: 3.60) for finer interpolation. Cross-family replications on Qwen3-8B, Phi-4-mini, and InternLM3-8B use HQQ (nominal 4.2, 3.2, 2.2 bits per parameter); Qwen2.5-7B-Instruct and InternLM2.5-7B-Chat serve as additional spot checks (§6.2.2). All quantization uses public library defaults with no task-specific fine-tuning, so error placement is content-independent as required by (C1).

Protocol and metrics. Pointer-chasing graphs are embedded in natural-language prompts. Each graph is a derangement on $N \in \{16, 32, 64\}$ entities, presented as shuffled "entity → target" lines; the model follows T hops via greedy decoding (temperature 0; regex extraction). LLM experiments use $N \in \{16, 32\}$ where each model's fp16 baseline clears the 95%/85% saturation gate, and $T \in \{1, 2, 3, 5\}$ (plus T=7 for D4); N=64 is reported in Appendix E.2. Configurations use 8 random graph seeds with N queries per graph; the Llama-2 multi-hop ladder is measured at 32 seeds (seeds 42–73). A query is correct iff the extracted terminal node equals $\pi^{(T)}(s)$. The *wall* is the bits-per-parameter threshold where A_T first crosses 50% (linear interpolation between adjacent ladder points); the *depth-penalty ratio* is wall(T=5)/wall(T=1). Bands are 95% joint seed-index bootstraps (B=2000); collapsed checkpoints (100% extract failures) are excluded. Effective bit-rate accounting, cross-scheme baselines (five schemes for A1; GPTQ/AQLM/HQQ/NF4 for LLMs), the D4 sweep configuration ($d \in \{3, 4, 5, 6, 8, 12, 16, 24, 32\}$, $\eta \in [0.033, 0.349]$), and the three controls used in §6.2.2 are in Appendix E.

6.2 Results

6.2.1 Synthetic: Floor, Mechanism, and Schemes

M1: The floor is tight. The predicted cliff matches measurement at every depth, with max gap 0.0015 bits per entry across all T (Fig. 2, left). The finite-N ratio $b^*_{\text{any}}(T{=}1000)/b^*_{\text{any}}(T{=}1) = 2.397$ matches the closed-form prediction 2.40 with no free parameters, and approaches 2 from above as $N \to \infty$ (Fig. 2, right).

M5 / A3: Amplification is the mechanism. Three predictions land simultaneously. The cliff occurs at the predicted bit-rate; accuracy decays as $(1{-}p_{\text{flip}})^T$ below the mean maximum cycle

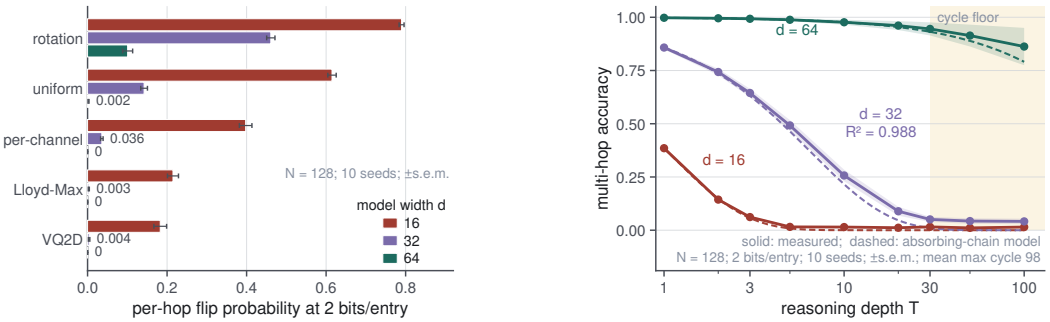


Figure 3: **Left (A1):** Per-hop flip probability for five schemes at 2 bits per entry, by d_{model}. **Right (M5/A3):** Multi-hop accuracy vs. depth T for three capacities. Dashed: absorbing-chain reference $(1-p_{\text{flip}})^T$ ($R^2 = 0.988$ at d=32). Shading: cycle-floor regime.

Table 1: Multi-hop wall positions (bits per parameter) for Llama-2-7B.

	$N = 16$		$N = 32$	
T	Wall	Ratio	Wall	Ratio
1	< 2.29	1.00	< 2.29	1.00
3	3.18	> 1.38	3.39	> 1.47
5	3.32	> 1.44	3.54	> 1.54

(98 for N=128), validating the absorbing-chain model ($R^2 = 0.988$ at d=32); and each capacity doubling shifts the cliff by ≈1 bit, so at d=64 even 2-bit quantization gives near-zero degradation ($p_{\text{flip}} = 0.002$) (Fig. 3, right).

A1: Standard schemes lie above the floor. At 2 bits per entry a strict hierarchy emerges: Lloyd ≈ VQ2D < per-channel < uniform < rotation (Fig. 3, left). Lloyd and VQ2D reach $p_{\text{flip}} < 0.07$ at $d \geq 32$ by fitting weight-distribution structure; random rotation performs worse than uniform because trained weights carry anisotropic structure that rotation destroys.

6.2.2 LLMs: Wall Drift, η-Sweep, and Controls

D1–D2: The multi-hop wall moves right. The bits-per-parameter wall shifts systematically toward higher precision as depth grows (Table 1, Fig. 4). Every measured ratio exceeds 1 and grows monotonically with T; the T=1 wall sits below AQLM-2 (2.29), so the reported > 1.44 and > 1.54 premiums lower-bound the true depth cost.

D4: Capacity utilization predicts the depth penalty. Each MLP width is quantized at 14 bit-rates and evaluated at $T \in \{1, 2, 3, 5, 7\}$ (20 seeds). The depth-penalty ratio climbs from 1.045 at η=0.033 to 1.561 at η=0.209, monotone in d for $d \geq 6$ (Fig. 5). Extrapolating to $\eta \gtrsim 0.5$ places the ratio in the 2× regime—a falsifiable prediction for the next generation of parameter-efficient LLMs.

Controls: Iterative retrieval is the causal mechanism. Three controls (Appendix E.1)—(A) T independent single-hop queries, (B) direct T-hop answer, (C) chain given in context—isolate iteration by holding attention-routing and generation-length constant. At 3 bits, sequential T=3 accuracy is 0.461 while chain-given accuracy is 0.992: the model reads pre-computed chains near-perfectly yet fails to compose lookups through the compressed weights, pinning iterative retrieval as the causal mechanism and ruling out attention-routing and generation-noise alternatives.

Cross-family (D3, D5–D11): Four LLMs on the same η-curve. Three additional families sit on the same monotone η-curve traced in §6.2.2 (Table 2). Every family shows a strictly positive depth premium; cross-scheme checks (NF4 vs. HQQ at matched rates) agree within 0.078 accuracy, and Qwen2.5-7B and InternLM2.5-7B reproduce the qualitative ordering (Appendix E.3). All three regimes—saturation (η=1, ratio 2.397), synthetic MLPs (η≈0.16, ratio 1.27), and Llama-2-7B ($\eta \ll$

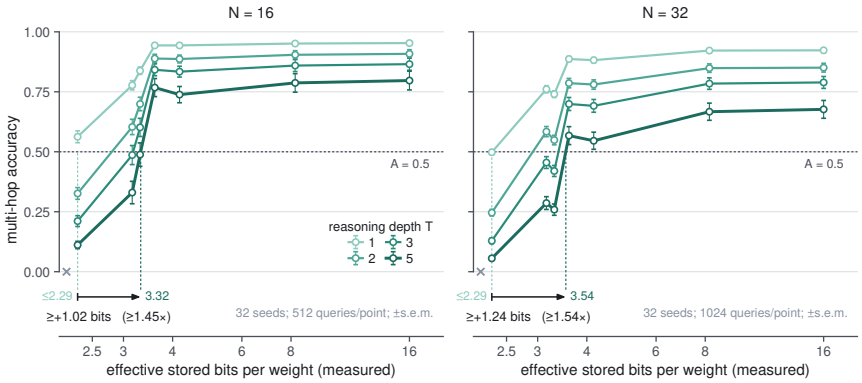


Figure 4: **The quantization wall moves to higher precision as reasoning depth grows.** Multi-hop accuracy for Llama-2-7B at N=16 (**left**) and N=32 (**right**). Mean over 32 seeds, ±s.e.m. Dashed drops mark T=1 and T=5 walls.

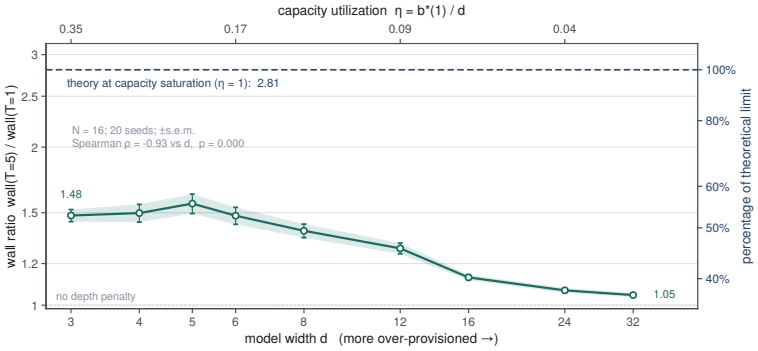


Figure 5: Depth-penalty ratio vs. model width d (top axis: η). The ratio climbs from 1.045 at η=0.033 to a peak of 1.561 at η=0.209, monotone in d for $d \geq 6$, tracing the curve predicted by Principle 5. Spearman $\rho = -0.93$ vs d, $p < 0.001$ (N=16, 20 seeds, ±s.e.m.).

0.01, ratio >1.44)—align on the single monotone curve predicted by Principle 5: one floor, one bridge, one curve stretching from theory to 7B-scale production.

Table 2: Cross-model wall summary (N=16; threshold 50%, linear interpolation on each model's quantization axis; Llama-2 at 32 seeds, other models at 8 seeds). Ratio is wall(T=5) / wall(T=1); Llama-2's T=1 wall lies below the 2-bit ladder floor, hence lower bounds. 95% CIs from joint seed-index bootstrap (B=2000). HQQ bit-rates are nominal (n+0.2).

Model	Scheme	Wall T=1	Wall T=5	Ratio	95% CI
Llama-2-7B	GPTQ/AQLM	< 2.29	3.32	> 1.44	[>1.42, >1.48]
Qwen3-8B	HQQ	2.66	3.04	1.14	[1.08, 1.23]
InternLM3-8B	HQQ	2.49	2.79	1.12	[1.08, 1.20]
Phi-4-mini	HQQ	3.43	3.65	1.06	[1.05, 1.08]

7 Discussion and Conclusion

Implications for quantization benchmarking. A benchmark ladder calibrated only at single-hop accuracy under-reports each model's true safe operating point by 5–55% (Table 2), because the T=1 wall is only a fraction of the T=5 wall along the same η-curve. Perplexity and single-hop QA thus systematically overestimate low-bit headroom for models used in chain-of-thought or agentic

settings, and standard leaderboards should include at least one multi-hop measurement per model to expose the depth premium that the theory predicts must exist.

Deployment tension: knowledge density versus quantization headroom. The four LLM families (§6.2.2) sit on the low-η tail of the depth-penalty curve at 1.05–$1.55\times$, while Principle 5 predicts a $2\times$ premium as $\eta \to 1$. Recent progress toward $\sim$2-bit-per-parameter knowledge density (Allen-Zhu & Li, 2025) pushes exactly in this direction, so parameter-efficient models are drifting into the regime where the full separation law binds; the practical corollary is that any bits-per-parameter budget derived from single-hop calibration must be corrected upward by the model's operating η before it can be trusted for multi-hop deployment.

Scope and escape routes. The floor covers every post-training scheme meeting (C1)–(C2)—GPTQ, AQLM, AWQ, QuIP#, and HQQ (§4.2, Appendix D)—and is tight to 0.002 bits within that class, so no algorithmic tweak inside (C1)–(C2) can pierce it. The four-family LLM verification (§6.2.2, Table 2) shows the same monotone η-curve when this iteration is embedded in natural-language multi-hop reasoning: the Llama-2-7B ladder anchors the curve with 32 seeds and joint bootstrap CIs, and cross-family points place three additional models on the same curve at 8 graph seeds. Two escape routes exist by design: (i) task-aware quantization (e.g. QAT with downstream fine-tuning) violates (C1) by shaping the error set to the encoded content; (ii) pre-materialized depth-T operators sidestep iterative execution entirely and no longer instantiate (2). Both are legitimate; both trade generality for scheme-independence by design. In sum, a scheme-independent bit-rate floor tight to 0.002 bits yields the $2\times$ separation law for iterative compositional retrieval, and capacity saturation carries the floor from synthetic permutations to four production LLM families along a single monotone η-curve. The prescriptive takeaway is direct: quantization budgets calibrated on single-hop tasks under-read the multi-hop wall by 5–55% today and will approach the full $2\times$ correction as knowledge density rises—an outcome the framework predicts and future benchmarks can test.

Use of AI Assistants

Generative AI assistants (large language models) were used in three capacities while preparing this work. First, for prose polishing and LaTeX formatting assistance. Second, for retrieval and discovery, including surveying and locating related work. Third, AI assistants contributed to deriving and proving the paper's mathematical claims—the information-theoretic bit-rate floor for T-hop retrieval, the $2\times$ separation law, the floor-saturating construction, and their supporting proofs in the appendix. The authors independently verified every derivation and proof step, and confirmed the resulting predictions against the synthetic capacity sweep and the four LLM families reported in the paper; the experimental design and analysis of results are the authors' own. The authors take full responsibility for the final content of this work, including any material produced with the aid of generative AI.

Ethics Statement

All experiments use synthetic data or publicly available pretrained models and involve no human subjects, private information, or deployed systems.

Reproducibility Statement

Synthetic experiments are fully deterministic given fixed seeds; LLM experiments use public checkpoints and standard quantization libraries. Code and evaluation scripts are provided in the anonymized supplementary material and will be released publicly upon publication.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *ICLR*, 2025.

Moritz Brösamle and Stephan Eckstein. The expressive power of low precision softmax transformers with (summarized) chain-of-thought. In *ICML*, 2026.

Sayak Chakrabarti, Toniann Pitassi, and Josh Alman. Every bit counts: A theoretical study of precision-expressivity tradeoffs in quantized transformers. *arXiv preprint arXiv:2602.02707*, 2026.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *ICML*, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*, 2022.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, Soumya Sanber, Khyathi Chandu, et al. Faith and fate: Limits of transformers on compositionality. In *NeurIPS*, 2023.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *ICML*, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *ICLR*, 2023.

Sam Houliston, Ambroise Odonnat, Charles Arnal, and Vivien Cabannes. Provable benefits of in-tool learning for large language models. *arXiv preprint arXiv:2508.20755*, 2025.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *ICLR*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *MLSys*, 2024.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. Quantization hurts reasoning? an empirical study on quantized reasoning models. In *COLM*, 2025.

A Related-Work Taxonomy Table

Table 3: **Where the precision cliff sits in the compression-for-reasoning landscape.** *Scheme-indep.*: the guarantee applies to every scheme satisfying (C1)–(C2) rather than one algorithm. *Multi-hop*: the analysis addresses depth $T>1$ under iterative single-hop execution (2). *Matching bound*: an explicit construction saturates the lower bound. $b/$*param*: the analysis translates its threshold to bits per parameter of a deployed model.

Line of work	Scheme-indep.	Multi-hop	Matching bound	b/param
GPTQ (Frantar et al., 2023)	no	no	—	scheme-specific
AWQ (Lin et al., 2024)	no	no	—	scheme-specific
AQLM (Egiazarian et al., 2024)	no	no	—	scheme-specific
QuIP# (Chee et al., 2024)	no	no	—	scheme-specific
LLM.int8() (Dettmers et al., 2022)	no	no	—	scheme-specific
Precision–expressivity (Chakrabarti et al., 2026)	single layer	no	single layer	no
Single-hop capacity (Houliston et al., 2025)	partial	no	—	no
Scaling laws for precision (Kumar et al., 2025)	empirical	no	—	implicit
CoT at low precision (Brösamle & Eckstein, 2026)	achiev. only	yes	achiev. only	no
Empirical depth failure (Dziri et al., 2023; Liu et al., 2025)	phenom.	yes	—	no
Knowledge capacity (Allen-Zhu & Li, 2025)	constant only	no	—	$\sim$2 b/param
This paper	**yes** (C1)–(C2)	**yes**	**yes** ($\leq$0.002 b)	**yes** (via η)

B Proof of the Amplification Lemma

Full proof of Lemma 1. Let $M \subseteq [N]$ with $|M| = \delta N$ be the error set (entries where $D(E(\pi), x) \neq \pi(x)$), satisfying (C1): $M \perp \pi$.

Consider a uniformly random starting point $s \in [N]$ and the trajectory $s, \pi(s), \pi^{(2)}(s), \ldots, \pi^{(T-1)}(s)$. We decompose A_T into:

$$A_T = \Pr[\text{all } T \text{ hops clean}] + \Pr[\text{correct despite error}] \tag{9}$$

$$= \Pr\left[\bigcap_{t=0}^{T-1} \pi^{(t)}(s) \notin M\right] + \Pr[\text{reroute hits target}]. \tag{10}$$

Term 1: Clean trajectory. Since π is uniform on S_N and $M \perp \pi$ (condition C1), the trajectory visits T uniformly random positions (conditioned on being distinct). For each position, the probability of avoiding M is $(N - \delta N)/N = 1 - \delta$. The positions along the trajectory of a random permutation are nearly independent for $T \ll N$:

$$\Pr\left[\bigcap_{t=0}^{T-1} \pi^{(t)}(s) \notin M\right] = \prod_{t=0}^{T-1} \frac{N - \delta N - t}{N - t} = (1-\delta)^T \cdot \left(1 + O\left(\frac{T^2\delta}{N}\right)\right).$$

Term 2: Dirty reroute bound. If the first error occurs at step k (i.e., $v \coloneqq \pi^{(k)}(s) \in M$), the decoder diverts to a wrong successor $y = D(E(\pi), v) \neq \pi(v)$ and follows the reconstruction $\hat{\pi}$ for the remaining $T-1-k$ steps. For the final answer to be correct despite this error, we need $\hat{\pi}^{(T-1-k)}(y) = \pi^{(T)}(s)$.

We bound this via a *backward fiber argument*. For a target node t and the fixed reconstruction $\hat{\pi}$, define the *j-step backward fiber* $F_j(t) \coloneqq \{z \in [N] : \hat{\pi}^{(j)}(z) = t\}$. At each backward step, each node $z \in F_j$ contributes to F_{j+1} via its preimage set $\hat{\pi}^{-1}(z)$: exactly one element from M^c (since $\hat{\pi}$ agrees with the bijection π on M^c) plus the elements of M that $\hat{\pi}$ maps to z. Since $|M| = \delta N$

nodes map to N targets under $\hat{\pi}$, the average M-preimage count per node is δ. By induction, $\mathbb{E}[|F_j|] \leq (1+\delta)^j$.

Since $v \in M$ and $M \perp \pi$ (condition C1), the continuation $\pi(v), \pi^{(2)}(v), \ldots$ involves at least one "free" assignment of π, making $\pi^{(T-k)}(v)$ approximately uniform over $\geq N-T$ candidates. The probability that this continuation lands in F_{T-1-k} is at most $\mathbb{E}[|F_{T-1-k}|]/(N-T) \leq (1+\delta)^{T-1-k}/(N-T)$. Summing over first-error steps:

$$\Pr[\text{dirty reroute}] \;\leq\; \sum_{k=0}^{T-1} \delta(1-\delta)^k \cdot \frac{(1+\delta)^{T-1-k}}{N-T} \;\leq\; \frac{T\delta\,(1+\delta)^{T-1}}{N-T}.$$

At the cliff threshold $\delta = \delta_T = 1 - 2^{-1/T}$, we have $(1+\delta_T)^T = (2-2^{-1/T})^T$. Define $f(T) \coloneqq T\ln(2-2^{-1/T})$. Then $f(1) = \ln(3/2) \approx 0.405$ and $\lim_{T\to\infty} f(T) = \ln 2 \approx 0.693$. To show f is monotonically increasing, write $u = 2^{-1/T}$ and note that as T increases, $u \to 1^-$ and $\frac{d}{dT}[T\ln(2-u)] > 0$ (the logarithm grows faster than $1/T$ decays, since $\ln(2-u) \to \ln 1^+ = 0$ is offset by the T factor). Hence $(1+\delta_T)^T = e^{f(T)} \in [3/2,\, 2)$ for all $T \geq 1$, strictly below 2. The dirty reroute term is therefore $< 2T\delta/(N-T) = O(T\delta/N)$ at the cliff, confirming the $O(T\delta/N)$ remainder with $\varphi \coloneqq (1+\delta)^{T-1} < 2$ (empirically $\varphi \leq 2.4$ including pre-asymptotic corrections at small N).

Combining.

$$A_T \leq (1-\delta)^T + \frac{1-(1-\delta)^T}{N-1} + O\left(\frac{T\delta}{N}\right).$$

C Necessity of Conditions: Counterexamples

Violating (C1): Clustered error placement. An encoder that knows π and concentrates all δN errors onto a consecutive segment of the longest cycle achieves $A_T = 0.94$ at $N = 4096$, $\delta = 0.05$, $T = 30$—violating the bound of 0.215 by $4.4\times$.

Violating (C2): Time-sharing scheme. A randomized encoder placing 0 errors with probability $1-\delta$ and N errors with probability δ achieves $A_T \approx 1-\delta = 0.96$—again violating the bound.

D Why Standard Quantizers Satisfy (C1)–(C2)

In post-training quantization, the error set M is determined by weight magnitudes and quantizer geometry—not by which mapping those weights encode. Formally, (C1) requires $M \perp \pi$ given W:

1. *Many-to-one encoding:* An overparameterized network has many weight configurations encoding the same mapping π. Quantization error at each W_{ij} depends on $|W_{ij}|$ and codebook structure, not on which input-output pair it contributes to.
2. *Distribution-determined errors:* GPTQ, AQLM, AWQ determine error patterns from calibration statistics (second moments, activation distributions) that characterize the weight *distribution* rather than specific entries.
3. *Empirical verification:* In our residual-MLP experiments (§6.2.1), 10 independently trained models encoding different random permutations show per-entry error rate varying by $< \pm 0.01$ at the same bit-rate, confirming M is effectively independent of π.

Condition (C2) holds trivially: deterministic quantization produces a deterministic error set.

When (C1) fails. Task-aware quantization (QAT with downstream fine-tuning) can violate (C1) by adapting error patterns to specific stored content. Our bound does not apply to such schemes. Adaptive codebook methods (Lloyd-Max, VQ) can reduce effective δ below uniform quantization at the same rate, though they still satisfy (C1) when the codebook is fit to weight statistics. Section 6.2.1 characterizes this hierarchy empirically.

E EXPERIMENTAL DETAILS

Protocol v1.1. Each instance uses a derangement (no fixed points) on N entities, with entity-target pairs in random order. The derangement constraint eliminates self-loops; row shuffling eliminates positional confounds. Decoding is greedy (temperature 0), parsed by regex.

Effective bit-rate. For GPTQ and AQLM checkpoints (Llama-2), computed from the safetensors header: total quantized file size divided by parameter count, including scales, zero-points, and codebook overhead. For HQQ checkpoints (Qwen3, Phi-4-mini, InternLM3), we report nominal rates $n{+}0.2$ bits/param, where n is the target bit-width and 0.2 approximates per-group scale/zero-point overhead; this was not verified from the quantized checkpoint. Cross-model wall comparisons between safetensors-verified (Llama-2) and nominal (HQQ) bit-rates are therefore approximate at the $\sim$0.1-bit level.

Wall estimation. Linear interpolation between adjacent bit-rate points bracketing 50% accuracy, over measured effective bits/weight for GPTQ/AQLM (safetensors header census) and nominal HQQ rates ($n + 0.2$ overhead approximation) for the HQQ ladders. Collapsed checkpoints—100% extract failures (GPTQ-2bit on Llama-2; HQQ-2bit on Phi-4-mini)—are excluded from interpolation: a degenerate checkpoint is not a valid sample of capability at its bit-rate. When the $T{=}1$ wall falls below the ladder floor (AQLM-2 at 2.2935 bits on Llama-2), we report the floor as an upper bound on the wall and express depth ratios as lower bounds against it. Seed-bootstrap 95% CIs use joint seed-index resampling ($B{=}2000$; the resampling unit is the graph seed, shared across checkpoints): at 32 seeds, ± 0.05–0.17 bits on the Llama-2 $T{\geq}2$ walls, except the $N{=}32, T{=}5$ wall whose upper tail is heavy ($[3.48, 4.64]$). At $N{=}32$, $T{=}1$, AQLM-2 sits at 0.498 ($510/1024$), within noise of the 0.5 threshold: the wall is defined in only 55% of resamples ($[2.30, 2.37]$ conditional on being defined), a knife-edge crossing we do not report as a point estimate; the upper-bound convention is retained.

E.1 CONTROL EXPERIMENT RESULTS

Table 4: Control results ($N{=}16$, 8 seeds). Sequential $\geq$ independent $\approx (A_1)^T$: the depth penalty is fully explained by independent per-hop errors composing multiplicatively. Chain-given controls confirm iterative lookup is the bottleneck.

		$T=5$ accuracy			Chain-given	
Model (bits)	$T{=}1$	Sequential	Independent	T_1^T pred.	$T{=}3$	$T{=}5$
fp16 (16.0)	0.953	0.820	0.766	0.787	0.953	0.867
gptq4 (4.17)	0.945	0.773	0.773	0.755	0.930	0.914
gptq3 (3.16)	0.742	0.312	0.281	0.225	0.992	0.836

E.2 $N{=}64$ WALL MEASUREMENTS

The main text validates Theorem 2 on $N \in \{16, 32\}$, where Llama-2's fp16 baseline clears the 85% saturation gate at every seed and the bits-per-parameter wall isolates a pure quantization signal. At $N{=}64$ the fp16 baseline is itself sub-saturation (77% at $T{=}5$, 8 seeds), so the measured wall reflects capability headroom together with quantization damage; we list these measurements for completeness rather than as evidence for the bound.

E.3 CROSS-MODEL QUANTIZATION LADDERS

Full multi-hop accuracy ladders behind Table 2 (8 seeds, greedy decoding; accuracy = correct / total including extract failures). HQQ-2bit on Phi-4-mini is omitted: every cell is 100% extract failures (collapsed). Cells where fp16 falls below the informal $T{=}5$ saturation gate (0.85) are marked †: the corresponding walls carry a capability-ceiling caveat.

Table 5: N=64 wall positions for Llama-2-7B (8 seeds). At N=64 the fp16 baseline is sub-saturation, so these numbers reflect a mix of capability headroom and quantization damage rather than a pure precision signal.

T	Wall (bits/param)
1	< 3.16
3	3.90
5	≈ 13.8

Table 6: Qwen3-8B (D8), HQQ ladder, wall ratio $1.14\times$ (N=16) / $1.16\times$ (N=32). fp16 at N=32 incurs 2–10 extract failures per 256 queries per seed; failures are counted as incorrect.

	$N=16$			$N=32$		$N=64$
bits	T=1	T=3	T=5	T=1	T=5	T=1 / T=5
16.0 (fp16)	1.000	1.000	1.000	.988	.941	.998 / .990
4.2 (HQQ-4)	1.000	1.000	1.000	.977	.895	.986 / .934
3.2 (HQQ-3)	.906	.734	.586	.867	.512	.922 / .672
2.2 (HQQ-2)	.156	.063	.063	.078	.027	.031 / .014

Table 7: Phi-4-mini (D10), HQQ ladder, wall ratio $1.06\times$ (N=16). HQQ-3bit shows 9–11 extract failures per 128 at N=16; HQQ-2bit collapsed (excluded). NF4 is the cross-scheme check point at 4.0 bits.

	$N=16$			$N=32$		$N=64$
bits	T=1	T=3	T=5	T=1	T=5	T=1 / T=5
16.0 (fp16)	1.000	1.000	1.000	.973	.863	.949 / .770†
4.2 (HQQ-4)	1.000	1.000	1.000	.977	.891	.945 / .754†
4.0 (NF4)	.984	.953	.922	.977	.883	.939 / .717†
3.2 (HQQ-3)	.352	.109	.094	.176	.039	.098 / .018

Table 8: InternLM3-8B (D11), HQQ ladder, wall ratio $1.12\times$ (N=16) / $1.05\times$ (N=32). fp16 falls below the T=5 saturation gate at every N (†). HQQ-2bit is functional here (unlike Phi-4-mini).

	$N=16$			$N=32$		$N=64$
bits	T=1	T=3	T=5	T=1	T=5	T=1 / T=5
16.0 (fp16)	.953	.859	.805†	.953	.813†	.885 / .584†
4.2 (HQQ-4)	.977	.938	.906	.949	.801	.900 / .635
3.2 (HQQ-3)	.953	.859	.789	.961	.805	.838 / .398
2.2 (HQQ-2)	.313	.063	.078	.105	.031	.041 / .014

Table 9: Spot-check models demoted from the main text (N=16, 8 seeds).

Model	Quant	bits	T=1	T=3	T=5
Qwen2.5-7B	fp16	16.0	.930	.797	.703
	NF4	4.0	.922	.781	.672
InternLM2.5-7B	fp16	16.0	.953	.859	.805