Graph-Parameterized Capacity for Latent Multi-Hop Recall: Usage, Compression, and Composition

Anonymous authors
Paper under double-blind review

Abstract

When can knowledge stored in model parameters replace bounded knowledge-graph traversal? We price latent multi-hop storage by graph structure rather than raw fact count. We define coverage-conditioned support usage H, macro-decomposition credit S, and realizable width $W_{\mathcal{D}}$, and give a geometric construction with dimension quadratic in $W_{\mathcal{D}}$ and logarithmic in code count and inverse slack. For witness-executing position-factorized memories, we prove a restricted entry-count lower bound matching the $H - S$ charge on branch-functional, pair-balanced closures under stated selector and representation assumptions. For modular memories, we prove an expected confabulation-plus-abstention floor on unique-witness crossings for weight-local, frame-nonaligning composition with a latent crossing interface, under receiving-cap overlap and path-separation assumptions. Empirically, 14 theory-matched families share a coarse rung index on an $H \log_2 |V|$-scaled ladder, while realized prefactors remain family-dependent; the distractor control does not identify the $\log_2 |V|$ factor. Support diversity best predicts thresholds. Macro credit reduces explicit parameters under slot-elastic allocation, while trained macro gains mainly reflect shorter live retrieval depth. A receiving-gate audit isolates frame mismatch. Pretrained LoRA thresholds remain bottom-censored, while CoT reaches the measured floor at $k = 4$ on 9/9 chain and 8/9 segmented cell-seeds.

1 Introduction

A language model that must generate text grounded in a body of knowledge faces a substrate choice: keep the knowledge *external* — a knowledge graph traversed at inference time, where the binding resource is depth and context — or *internalize* it into parameters via fine-tuning, adapters (Biderman et al., 2024), sparse memory slots (Lin et al., 2025), or trained KV objects (Eyuboglu et al., 2026), where the binding resource is capacity. Theory has begun to price the second substrate — ∼2 bits per parameter (Allen-Zhu & Li, 2024), near-optimal associative storage (Nichani et al., 2025), a capacity–depth tradeoff for latent k-hop recall that chain-of-thought collapses (Ravfogel et al., 2026; Johnston & Belrose, 2025; Chen, 2025) — but every such result is proved over *unstructured* fact sets, and Ravfogel et al. (2026) leave the structured case open. Real corpora, clinical knowledge graphs above all, have hubs of enormous degree, heavily reused sub-paths, and typed schemas. For explicit entry-based memories, path collapse and cross-query reuse can reduce storage charge while hubs can increase support usage; whether gradient-trained continuous memories realize those same combinatorial savings is a separate empirical question. Fact-count laws cannot separate these effects. Two adjacent questions have the same shape. If knowledge is internalized modularly, independently trained memories collapse to near chance when naively composed (Hardalov et al., 2026; Balesni et al., 2025), and no existing theory predicts composability from the *data*. And when a system should internalize at all — retrieval dominates fine-tuning on low-frequency and novel multi-hop knowledge (Soudani et al., 2024; Yang et al., 2026) — has no formal account.

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.

Our paper is that these are one question: a *sufficiency function* mapping a corpus graph $\mathcal{G}$, a query class with distribution μ, and a tolerance to the minimum resource a substrate needs, with **the graph setting the price** in either currency. This paper establishes it for the parametric substrate, and measures where the price is paid and where it is not. Related work is summarized in Section 2; the formal scope and empirical status of our claims, together with notation, are collected in Appendix C. Building on this perspective, we next summarize the paper's main theoretical, empirical, and mechanistic contributions.

1. **Graph-parameterized capacity.** Theorem 1 gives, for every decomposition $\mathcal{D}$, a full parameter bound with $d_{\mathcal{D}} = O(W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta))$; on branch-functional, workload-dominant closures it reduces to the $H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$ charge. Theorem 2 gives a restricted matching entry-count lower bound for witness-executing position-factorized memories under pair-balance, selector-optimality, and explicit-entry assumptions; the unrestricted transformer extension remains Conjecture 1.
2. **Composition.** Theorem 3 gives an expected confabulation-plus-abstention floor for frame-nonaligning composition of independently coded, path-separating gated memories on unique-witness crossings, under an explicit receiving-cap overlap assumption.
3. **Measured structure effects.** On the theory-matched gated substrate, all 14 families share the same coarse rung index on an $H \log_2 |V|$-scaled ladder, while realized prefactor brackets remain family-dependent; their seven-family intersection envelope is $c_f \in (0.578, 1.228]$. The distractor control resolves no rung shift as unused vocabulary grows and therefore does not identify the $\log_2 |V|$ factor. Support diversity predicts threshold better than usage concentration. Explicit macro-decomposition savings depend on allocation shape, while the trained macro gain is explained by reduced live retrieval depth rather than measurable cross-query reuse.
4. **Mechanism audits and negative results.** A receiving-gate alignment audit isolates the receiving-gate mechanism; forced-width tests are consistent with the theorem's $W_{\mathcal{D}}^2$ dependence under fixed-C trajectories; pretrained LoRA thresholds remain bottom-censored through 10^4-edge closures; and CoT reaches the measured floor for segmented $k = 3$ and for 9/9 chain versus 8/9 segmented $k = 4$ cell-seeds under their stated criteria.

2 Related Work

Capacity results for transformers price unstructured facts (Allen-Zhu & Li, 2024; Cabannes et al., 2024; Nichani et al., 2025; Ravfogel et al., 2026; Chen, 2025); Sanford et al. (2024) study in-context graph algorithms. Mechanistic work directly probes factual recall and shortcut-controlled latent multi-hop reasoning (Geva et al., 2023; Yang et al., 2024; 2025); layer-order inversion, back attention, and controlled data/depth scaling show that latent computation need not follow a simple hop-aligned execution picture and can be optimization-limited (Liu et al., 2026; Yu et al., 2025; Yao et al., 2025). Modular-memory and merging work conditions primarily on weights rather than graph structure (Eyuboglu et al., 2026; Hardalov et al., 2026; Lin et al., 2025; Ilharco et al., 2022; Ortiz-Jimenez et al., 2023; Yadav et al., 2023; Song & Zheng, 2026). On retrieval versus internalization, Pulici et al. (2026) compile a KG into per-entity adapters and find storage without similarity-retrievability, and structured multi-hop RAG (Huang et al., 2026; Dang et al., 2026; Lien et al., 2026; Lin et al., 2026; Hamid et al., 2025; Ghosh et al., 2024) populates the traversal side. Appendix C summarizes the formal scope and empirical status of the paper's principal claims.

3 A Theory of Structural Sufficiency

3.1 Setup

A *corpus graph* is a typed multigraph $\mathcal{G} = (V, E, R)$ with facts $E \subseteq V \times R \times V$; a k-hop chain query $q = (v_0, r_1, \ldots, r_j)$, $j \leq k$, has answer set $\mathrm{ans}(q)$ (endpoints of paths from v_0 realizing $r_1 \cdots r_j$) and any such path is a *witness*; $\mathcal{Q}_k(\mathcal{G})$ is the set of answerable queries with distribution μ, and an output is *correct* if it lies in $\mathrm{ans}(q)$, so answer multiplicity needs no uniqueness assumption.

Definition 1 (ε-faithfulness at coverage γ)**.** *A memory system M is ε-faithful at coverage γ on $(\mathcal{Q}_k, \mu)$ if*

$$\Pr_{q\sim\mu}\big[\, M(q) \neq \bot \ \wedge\ M(q) \notin \mathrm{ans}(q)\,\big] \le \varepsilon \qquad \text{and} \qquad \Pr_{q\sim\mu}\big[\, M(q) = \bot\,\big] \le 1-\gamma. \tag{1}$$

Faithfulness penalizes confabulation, never refusal; coverage is reported jointly so that faithfulness cannot be bought by abstaining everywhere. All bounds below fix a target pair (ε, γ) and use the single slack symbol $\eta := \min(\varepsilon,\, 1-\gamma)$ throughout.

For substrate $\sigma \in \{\mathsf{param}, \mathsf{trav}\}$ with resource ρ (parameters; depth and probe budget), $\mathsf{Suff}_\sigma(\mathcal{G}, \mathcal{Q}_k, \mu, \varepsilon, \gamma)$ is the least ρ at which some system of that substrate is ε-faithful at coverage γ.

Definition 2 (γ-retained closure; coverage-conditioned support usage)**.** *Fix (μ, γ) and a deterministic reference selector w_{ref}; all ties are resolved by fixed deterministic query/witness orders. Let $\mathcal{C}_k^\mu(\gamma)$ be a minimum-cardinality query set of μ-mass at least γ. Define*

$$\mathsf{use}_k(f,i) = \mathbf{1}[\exists q \in \mathcal{C}_k^\mu(\gamma) : f \text{ is hop } i \text{ of } w_{\mathrm{ref}}(q)], \qquad H_k = \sum_{f\in E}\sum_{i=1}^{k} \mathsf{use}_k(f,i). \tag{2}$$

Thus H_k counts distinct (fact, hop-position) pairs touched by retained reference witnesses; we write $H_k^{\mu,\gamma}$ at fixed $(\mu, \gamma, w_{\mathrm{ref}})$. The tolerance ε enters faithfulness and lower-bound slack, not this support definition.

Definition 3 (Macro decompositions and decomposition credit)**.** *A macro is a relation string m together with retained fact-paths realizing it. An admissible $\mathcal{D}$ rewrites each $w_{\mathrm{ref}}(q)$ as at most k macros/single facts; let $K_{\mathcal{D}}$ be the maximum resulting depth. With ℓ a primitive relation or macro label,*

$$H_k^{\mathcal{D}} = \#\{((u,\ell,v), i) : (u,\ell,v) \text{ occurs at slot } i \text{ in a retained } \mathcal{D}\text{-rewritten witness}\}.$$

Macro labels are charged separately even when endpoints coincide, and $S_k^{\mu,\gamma} := H_k^{\mu,\gamma} - \min_{\mathcal{D}} H_k^{\mathcal{D}} \ge 0$. For fixed $\mathcal{D}$, making macro identities query-private defines $H_k^{\mathrm{priv},\mathcal{D}}$ and the exact accounting split

$$H_k^{\mu,\gamma} - H_k^{\mathcal{D}} = (H_k^{\mu,\gamma} - H_k^{\mathrm{priv},\mathcal{D}}) + (H_k^{\mathrm{priv},\mathcal{D}} - H_k^{\mathcal{D}}) =: S_{\mathrm{collapse}}(\mathcal{D}) + S_{\mathrm{reuse}}(\mathcal{D}). \tag{3}$$

The split is accounting, not a nonnegativity claim. We make no worst-case approximation-ratio claim for the greedy decomposition used in experiments; every admissible $\mathcal{D}$ still yields a valid upper bound.

Hubs raise $H_k^{\mu,\gamma}$ (a fact is charged per position it serves); macro decomposition can lower $H_k^{\mathcal{D}}$ either by collapsing several hops within a query or by reusing a macro across queries; and parallel branches under one (head, relation) pair inflate the width $W_{\mathcal{D}}$ below. Table 3 (Appendix A) works these effects on toy graphs, and Remarks 2 and 4 give the support and greedy–optimal details.

3.2 Upper bound

The dimension bound below is governed by the *realizable width* $W_{\mathcal{D}}$ of a decomposition's entry table (Definition F.2.1): the maximum, over table-realizable macro strings and hop prefixes, of the number of fact-paths from one head realizing that prefix through the retained branch sets — a path count, not a degree; it is 1 on branch-functional closures and exceeds 1 exactly when parallel branches under one (head, relation) pair serve diverging retained queries.

Definition 4 (Model class $\mathcal{M}_{\mathrm{geo}}(k,d)$)**.** *A query executes up to k gated key–value blocks, with block i computing $x_i = \sum_{j\in T_i} v_j \sigma(\langle p_i, \kappa_j\rangle - b)$ and halting after its $K_{\mathcal{D}}(q)$ active slots. The head decodes the final state to a retained candidate endpoint or $\bot$. Counting table entries, head/value/relation/macro/slot codes, and shared maps gives*

$$\Theta\left(d\sum_i |T_i|\right) + \Theta(d(|V_{\mathrm{heads}}| + |V_{\mathrm{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}})) + O(d^2). \tag{4}$$

Dictionary serialization and query-to-macro parsing are separate resources (Remark 5).

Theorem 1 (Graph-parameterized capacity, upper bound). *Let $W_{\mathcal{D}}$ be the realizable width of the entry table induced by $\mathcal{D}$ (Definition F.2.1), let $K_{\mathcal{D}}$ be its effective macro depth, and write*

$$N_{\mathcal{D}} \coloneqq H_k^{\mathcal{D}} + |V_{\text{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}} + 1, \qquad d_{\mathcal{D}} \coloneqq C W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta). \tag{5}$$

for a sufficiently large universal constant C. Because $|V_{\text{values}}(\mathcal{D})| \le H_k^{\mathcal{D}}$, the definition of $N_{\mathcal{D}}$ already dominates the candidate-output code count. For every admissible $\mathcal{D}$ there is a configuration in $\mathcal{M}_{\text{geo}}(k, d_{\mathcal{D}})$ that is ε-faithful at coverage γ on $(\mathcal{Q}_k, \mu)$ and uses

$$P_{\mathcal{D}} = O\big(d_{\mathcal{D}}^2 + \big(H_k^{\mathcal{D}} + |V_{\text{heads}}| + |V_{\text{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}\big) d_{\mathcal{D}}\big)\,. \tag{6}$$

real-valued memory parameters. Consequently the unconditional parameter sufficiency bound is

$$\mathsf{Suff}_{\mathit{param}} = O\Big(\min_{\mathcal{D}} \big[d_{\mathcal{D}}^2 + \big(H_k^{\mathcal{D}} + |V_{\text{heads}}| + |V_{\text{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}\big) d_{\mathcal{D}}\big]\Big)\,.$$

If a particular $\mathcal{D}$ additionally satisfies the workload-dominance *conditions $H_k^{\mathcal{D}} \ge d_{\mathcal{D}}$ and $H_k^{\mathcal{D}} \ge |V_{\text{heads}}| + |V_{\text{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}$, then $P_{\mathcal{D}} = O(H_k^{\mathcal{D}} d_{\mathcal{D}})$. Thus, only under this condition may the bound be simplified to $O(H_k^{\mathcal{D}} W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta))$. On a branch-functional closure ($W_{\mathcal{D}} = 1$), if a charge-minimizing decomposition also satisfies workload dominance, this gives*

$$P_{\mathcal{D}^*} = O((H_k^{\mu,\gamma} - S_k^{\mu,\gamma}) \log(N_{\mathcal{D}^*}/\eta))\,, \qquad \mathcal{D}^* \in \arg\min_{\mathcal{D}} H_k^{\mathcal{D}}. \tag{7}$$

The theorem's primary statement is the full minimization above; the $H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$ display is its branch-functional, workload-dominant specialization. The construction additionally requires a serialized macro dictionary and parsing work, accounted separately in Remark 5.

By the dimension requirement in Eq. (5), the realizable width enters quadratically through $W_{\mathcal{D}}^2$, while the remaining dependence is logarithmic in the code count and inverse slack. The proof (Appendix F) draws i.i.d. spherical codes, merges parallel branches into one entry per (head, macro, active slot), and controls every gate by a single pairwise-incoherence event. Queries execute their $K_{\mathcal{D}}(q)$ active macro slots and halt. Macro dictionary storage and $O(k)$ parsing work are reported separately from the real-valued memory parameters (Remark 5, Appendix A).

3.3 Information ceiling and lower bound

Proposition 1 (Information ceiling). *$\mathcal{G}$ is describable in $O(|E|(\log|V| + \log|R|))$ bits. For the CoT claim, assume that the position-free set E_{used} of primitive facts used by the retained reference witnesses is functional: $|\{v : (u, r, v) \in E_{\text{used}}\}| \le 1$ for every (u, r). With intermediate emission, iterated single-hop lookup then achieves the target with $O(|E_{\text{used}}| \log|E|)$ parameters (Appendix F, F.4). The graph-description ceiling holds without this assumption; the stated CoT construction uses it. Define*

$$\text{LatentPremium} := \max\{0, (H_k^{\mu,\gamma} - S_k^{\mu,\gamma}) - |E_{\text{used}}|\}.$$

Because $H_k^{\mu,\gamma} - S_k^{\mu,\gamma} \le H_k^{\mu,\gamma} \le k|E_{\text{used}}|$, this premium is at most $(k-1)|E_{\text{used}}|$ for fixed k. Under the position-free functionality assumption, CoT removes the positive position-sensitive entry-count premium by re-tokenizing intermediate vertices and reusing one position-free one-hop table. Extending this guarantee to branching one-hop tables requires an additional branch-selection or search argument.

Definition 5 (Position-factorized class $\mathcal{F}_{\text{pf}}$). *A latent system is position-factorized if hop i factors through a discrete table T_i such that (a) resolving a witness at slot i requires a matching entry, (b) one entry resolves at most one labeled macro-fact per slot, and (c) for every $q \in \mathcal{Q}_k(\mathcal{G})$ with $M(q) \ne \perp$ and $M(q) \in \text{ans}(q)$, the execution of $M(q)$ resolves every slot of a realizable witness ending at $M(q)$ using the corresponding tables. Condition (c) applies to all correctly answered queries, including any outside the fixed retained closure. Condition (c) is an explicit witness-execution requirement: systems that produce a correct endpoint directly from a head-to-answer association without resolving a witness are outside the class. The associative constructions of Nichani et al. (2025); Ravfogel et al. (2026) fit this template; Theorem 1's construction does so on branch-functional closures, while its $W_{\mathcal{D}} > 1$ merged entries leave the class (Appendix F).*

Theorem 2 (Lower bound in $\mathcal{F}_{\mathrm{pf}}$). *Set $\varepsilon' := \varepsilon + (1-\gamma)$ and let*

$$\widetilde{H}_k^{\varepsilon'} := \min_{w(\cdot)} \min_{\substack{C \subseteq \mathcal{Q}_k(\mathcal{G}) \\ \mu(C) \geq 1-\varepsilon'}} \min_{\mathcal{D}} H_k^{\mathcal{D}}(C, w). \tag{8}$$

the decomposition charge minimized over witness selectors, over $(1-\varepsilon')$-mass sub-closures, and over admissible decompositions. (i) Any $M \in \mathcal{F}_{\mathrm{pf}}$ (Definition 5) that is ε-faithful at coverage γ on $(\mathcal{Q}_k, \mu)$ satisfies $\sum_i |T_i| \geq \widetilde{H}_k^{\varepsilon'}$. (ii) If moreover μ is β-pair-balanced (Definition 8, Appendix A: removing τ query mass deletes at most a $\beta\tau$ fraction of distinct pairs), then $\widetilde{H}_k^{\varepsilon'} \geq (1-\beta\varepsilon') \min_{w(\cdot)} \min_{\mathcal{D}} H_k^{\mathcal{D}}$. (iii) If additionally the fixed reference selector w_{ref} of Definition 2 attains the selector minimum, $\min_{w(\cdot)} \min_D H_k^D = \min_D H_k^D(w_{\mathrm{ref}})$ — an explicit hypothesis, not established in general — then, under the explicit-entry representation model, the parameter count P satisfies $P \geq 2(1-\beta\epsilon')(H_k^{\mu,\gamma} - S_k^{\mu,\gamma})d_{\min}$. In particular, if $\beta\epsilon' \leq 1/2$, then $P \geq (H_k^{\mu,\gamma} - S_k^{\mu,\gamma})d_{\min}$. Each table entry is stored as a key–value pair of scalar vectors of dimension $d_{\min}$ and therefore uses at least $2d_{\min}$ scalars.

Theorem 2 is fundamentally an entry-count result. Identifying its selector-minimized charge with the reference-selector $H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$ requires the additional assumption $\min_w \min_{\mathcal{D}} H_k^{\mathcal{D}} = \min_{\mathcal{D}} H_k^{\mathcal{D}}(w_{\mathrm{ref}})$. Part (iii) converts this charge to parameters only in the explicit-entry representation model and with $\beta\varepsilon' \leq 1/2$; shared-weight or superposed encodings can fall outside that conversion. The synthetic audit has unique witnesses and therefore no nontrivial selector choice. Exact toy checks establish optimality only for those toy instances, while the clinical comparisons cover only the tested heuristics; neither establishes global selector optimality on general workloads (Appendix A).

Conjecture 1 (General constant-depth lower bound). *Theorem 2 extends to all constant-depth transformers without intermediate emission. This remains open: the synthetic and clinical pretrained-capacity experiments are bottom-censored, the preregistered real-KG replication onset is optimization-limited, and no standard multi-hop transformer arm yields an uncensored capacity-threshold test (Section 4). Selector optimality in Theorem 2(iii) is a separate hypothesis.*

The matching claim is scoped to branch-functional closures (Remark 6) and assumes β-pair-balancedness as defined in Definition 8. The greedy kill-fraction audit at $\tau = 0.05$ is a diagnostic only: it does not establish the required inequality for every $\tau \in (0, 1]$, every admissible witness selector and decomposition, and every removable query set of mass at most τ. The full pair-balancedness condition therefore remains an explicit hypothesis of Theorem 2(ii)–(iii), rather than a conclusion of this audit (Remark 7). Proofs and the transformer mapping are in Appendix A.

3.4 Composition

Partition $E = E_1 \cup \cdots \cup E_m$; train memories $M_1, \ldots, M_m$ independently, one per cell, with a common frozen base model (token embeddings shared; all task-trained latent codes private). Let $\mathrm{cross}_k(\mathcal{P})$ be the retained μ-mass of queries whose reference witness uses facts from more than one cell; this is the empirical crossing statistic reported in the modular-composition experiment and receiving-gate alignment audit. For the formal composition floor, define the *unavoidable unique-witness crossing mass*

$$\mathrm{cross}_k^{\mathrm{uniq}}(\mathcal{P}) := \mu\Big(\big\{q \in \mathcal{C}_k^{\mu}(\gamma) : q \text{ has a unique realizable witness, and that witness uses facts from more than one cell}\big\}\Big).$$

Definition 6 (Weight-local and frame-nonaligning composition). Φ *is* weight-local *if it uses only $(M_1, \ldots, M_m)$ and the frozen base, with no data or gradient steps. It is* frame-nonaligning *if, for every private isometry O_i, rotating cell i acts only on structures inherited from that cell,*

$$\Phi(M_1, \ldots, O_i M_i, \ldots, M_m) = \rho_i(O_i)\Phi(M_1, \ldots, M_i, \ldots, M_m),$$

where $\rho_i(O_i)$ neither rotates another cell into i's frame nor estimates/undoes O_i from cross-cell coordinates. Slot union/KV concatenation qualify; a data-free operator that estimates a relative frame is weight-local but not frame-nonaligning.

A weight-local Φ has a *latent crossing interface* (Definition 7, Appendix A) if the composed probe at every retained crossing is a fixed function of the latent code M_A emits for the bridge vertex and of B-side structures — the bridge is not re-emitted as tokens; merging, slot union and KV concatenation qualify, CoT-style re-tokenization does not.

Assumption 1 (Latent codebook independence / receiving-cap overlap)**.** *For a retained unique-witness crossing $A \to B$, condition on the unrotated memories and frame-invariant side information. Draw an independent Haar relative frame O for A, so the emitted bridge code becomes $O\xi_A(v)$ while B's receiving region Cap^+ stays fixed. Assume*

$$\Pr_O[O\xi_A(v) \in \mathrm{Cap}^+ \mid M_A, M_B, \ \textit{side information}] \le p_{\mathrm{ch}}(d) + \delta_{\mathrm{align}}. \tag{9}$$

Here p_{ch} is the random-bridge cap probability and δ_{align} excess alignment. For branch-functional non-superposed bridges with no alignment channel, Remark 15 gives $p_{\mathrm{ch}} \le e^{-\Omega(d)}$; general superposed states retain the explicit cap-overlap quantity. Rotationally equivariant from-scratch training is one sufficient random-frame model, whereas cells sharing a pre-trained base need not satisfy it. The receiving-gate alignment audit measures this event directly.

A *path-separating gated-lookup memory* (Definition 9, Appendix A) uses the gated lookup of Definition 4, makes later hops functions only of the current state, and cannot reconverge from an off-witness crossing to $\mathrm{ans}(q)$. Slot union/KV concatenation of our $\mathcal{M}_{\mathrm{geo}}$ cells satisfy this condition; merged LoRA need not.

Theorem 3 (Expected composition floor on unique-witness crossings)**.** *Under Assumption 1, for every weight-local, frame-nonaligning Φ (Definition 6) with a latent crossing interface (Definition 7) whose composed system is a path-separating gated-lookup memory (Definition 9), write $\mathrm{err}_{\mathrm{conf}}(\Phi)$ for the μ-mass of retained queries the composed system answers incorrectly (confabulation) and $\mathrm{abst}(\Phi)$ for the μ-mass it routes to $\perp$. Let $\mathbf{O}$ denote the independent relative-frame draw(s) specified by Assumption 1, and write $\Phi_{\mathbf{O}}$ for the resulting composed system under those frames. Then*

$$\mathbb{E}_{\mathbf{O}}[\mathrm{err}_{\mathrm{conf}}(\Phi_{\mathbf{O}}) + \mathrm{abst}(\Phi_{\mathbf{O}})] \ \ge \ \big(1 - p_{\mathrm{ch}}(d) - \delta_{\mathrm{align}}\big) \cdot \mathrm{cross}_k^{\mathrm{uniq}}(\mathcal{P}). \tag{10}$$

The theorem bounds only the expected sum of confabulation and abstention in Eq. (10), not either term separately; proof is in Appendix G.

Theorem 3 is conditional on unique-witness crossings, weight-local and frame-nonaligning composition, the receiving-cap overlap assumption, a latent crossing interface, and path separation. It bounds the expected sum of confabulation and abstention over the stated relative-frame draws; it is not a universal impossibility result for independently trained models. Without path separation, the argument gives only the correct crossing-gate activation bound of Corollary 3, not an endpoint-error floor. Re-tokenization, co-occurrence, and shared-base alignment can remove or violate different premises (Appendix A). The cost-model and verifier discussion is deferred to Appendix B.

4 Pre-Registered Experimental Design and Results

Every experiment was pre-registered with kill criteria, 3–5 seeds, matched optimizer budgets, and coverage aligned across arms by tuning abstention to a fixed γ before comparing ε; thresholds are measured on the retained closure (the storage object of Theorem 1), so every (ε, γ) is in-sample by design. The study spans LoRA on Qwen2.5-0.5B for synthetic, clinical, and real-KG capacity probes; explicit constructions including forced-width and allocation controls; theory-matched trained gated memories for capacity scaling and macro-compression analyses; and modular-composition and receiving-alignment tests. Every deviation from a frozen protocol is a logged amendment; nothing was re-run to move a number (Appendix J.1).

Table 1: Receiving-gate alignment audit. The independent-frame control uses independent codebooks; the shared-frame control uses a shared frame. Both exact controls are receiving-complete; the trained union is shown only as a failed sanity control.

arm	subset	support	correct gate (given support)	random bridge	final correct
independent-frame	crossing	1.00	0.038–0.051	0.040–0.044	0.022–0.027
independent-frame	within	1.00	0.82–0.85	0.043–0.046	0.78–0.81
shared-frame	crossing	1.00	0.92–0.93	0.02–0.06	0.89–0.90
shared-frame	within	1.00	0.92–0.93	0.04–0.06	0.90
trained union	crossing	0.20–0.31	0.58–0.66	0.59–0.65	0.000
trained union	within	0.37–0.48	0.82–0.84	0.66–0.67	0.25–0.27

The two original pre-registered pretrained-capacity programs, on synthetic families and clinical KG subgraphs, ran to completion with every grid drained; every result file embeds its configuration hash and commit, and the reported results are generated from archived JSON by version-controlled analysis scripts (Appendix E).

Pretrained capacity is bottom-censored

All 215 synthetic k-hop-trained LoRA threshold runs sit at the grid floor $r = 1$ with $\varepsilon < 0.05$ at $\gamma = 0.9$; the clinical-KG rows are likewise floor-censored (Table 7). The α-sweep changes H_k^{greedy} by at most 1.75×, below the preregistered 2× kill threshold, so the pretrained capacity test is censored rather than passed or failed (Appendix D). The explicit construction has entries/$H_k^{\text{greedy}} = 1.00$ on the synthetic families and 0.996 on the clinical knowledge graph. In the preregistered forced-width follow-up, $W = 8$ fails through $d = 8192$ on 5/5 seeds and passes at $d = 16384 = 256W^2$; across $W \in \{2, 4, 8\}$, the tested fixed-normalization trajectories bracket $p \in (5/3, 2]$, while a free (C, p) law remains unidentified (Appendix D).

Modular composition and its escape mechanisms

The modular-composition experiment shows unit crossing failure for the independently coded $\mathcal{M}_{\text{geo}}$ slot union at every nonzero empirical crossing mass. The archive-verified $\mathcal{M}_{\text{geo}}$ joint-training and CoT controls reduce crossing failure to roughly 0.10–0.14 and, in the published split data, pay that residual as abstention (Table 8, Appendix D). Legacy LoRA composition rows are omitted because the current archive render does not reproduce their previously reported values; no theorem claim depends on them. A receiving-gate alignment audit then tests the theorem's receiving event with receiving-complete exact controls: independent-frame correct crossing-gate activation is 0.038–0.051, matching the random-bridge baseline 0.040–0.044, whereas shared-frame rises to 0.92–0.93 with final correctness 0.89–0.90. The archived trained union is not used as the mechanism test because receiving support is incomplete.

Interpretation. The two exact-oracle controls isolate the mechanism assumed by the composition theorem. Holding receiving support at one, independent latent frames reduce correct crossing-gate activation to the random-bridge rate while preserving high within-cell activation; sharing the frame restores both the gate and the final answer. This separates a representation-alignment failure from missing support or ordinary within-cell memorization. The trained-union rows fail that sanity control because receiving support is incomplete and random bridges already open a nominally correct gate too often; consequently they remain descriptive and are not used to estimate δ_{align}. Together with the modular-composition results, the audit supports the theorem's qualitative distinction between independently coded latent composition and shared-frame or re-tokenized escapes without treating the empirical crossing statistic as a plug-in estimate of the formal cap probability.

Real-KG replication and optimization control. The apparent real-KG rank-2 onset is single-hop and optimization-limited. On fresh MetaQA seeds it reproduces at baseline

Table 2: Realized prefactor brackets. Entries are (last failing, first passing] for $P/(H \log_2 |V|)$; "combined" intersects the coarse and fine sweeps.

family	coarse	fine	combined
`load_mid`	$(0.846, 1.228]$	$(1.020, 1.254]$	$(1.020, 1.228]$
`hubs3`	$(0.615, 1.223]$	$(1.027, 1.252]$	$(1.027, 1.223]$
`load_mid3`	$(0.640, 1.216]$	$(0.843, 1.025]$	$(0.843, 1.025]$
`hubs_reuse`	$(0.556, 1.218]$	$(0.706, 0.856]$	$(0.706, 0.856]$
`reuse_hi`	$(0.564, 1.215]$	$(0.701, 0.855]$	$(0.701, 0.855]$
`load_hi`	$(0.553, 1.216]$	$(0.582, 0.705]$	$(0.582, 0.705]$
`reuse_hi3`	$(0.557, 1.213]$	$(0.578, 0.704]$	$(0.578, 0.704]$

B50 on 3/3 seeds near $H \approx 8.1 \times 10^3$ and returns to rank 1 on 3/3 at B150. CWQ-LCC has no matched high-H baseline-B50 1×-step row in the preregistered replication; its saturated high-H B50 row uses 2× steps and is already rank 1. These runs therefore do not test multi-hop capacity on the real KGs (Appendix I).

Capacity scaling on theory-matched substrates

The theory-matched capacity-scaling experiment compares a separately trained, macro-free member of $\mathcal{F}_{\text{pf}}$—one position-specific table per hop, with no cross-position super-factorization—with two from-scratch transformer controls on 14 families ($k \in \{2, 3\}$, 5 seeds, budgets B and $3B$). This arm is distinct from the earlier pretrained $\mathcal{M}_{\text{geo}}$ capacity grid discussed in Appendix D, whose super-factorized code path lies outside $\mathcal{F}_{\text{pf}}$. The position-factorized gated-memory arm is uncensored on all families: all 70 coarse-grid cells first pass at the same nominal rung $c = 1.219$ on the approximately $\times 2.21$ $H \log_2 |V|$-scaled ladder, with identical $B/3B$ thresholds on 14/14 families. The realized first-pass ratio spans 1.196–1.242 because integer table width and fixed parameter terms perturb the nominal target; this is rung-index constancy rather than a universal point prefactor. For the same reason, realized last-fail/first-pass ratios need not preserve the nominal ladder spacing (for example, `load_mid` has $(0.846, 1.228]$). Log–log slopes of P^* versus H are 1.01 ($k = 2$) and 1.00 ($k = 3$).

The distractor control holds H bit-identical while $|E|$ grows 2k → 8k → 32k. All 15 cells (three sizes × five seeds) fail at the same architectural floor $P = 8450$ and first pass at the same coarse rung index, giving overlapping brackets $(8450, 12290]$, $(8450, 15618]$, and $(8450, 18178]$. Thus this control resolves no capacity-rung shift under a 16× expansion of never-queried edges; it is consistent with, but does not identify, a $\log_2 |V|$ factor. The normalized first-pass targets 1151/1135/1140 are descriptive because the ladder itself is scaled by $H \log_2 |V|$ (Figure 1). A fine ladder gives narrower intervals on all seven audited families; for `load_mid` and `hubs3`, however, the coarse pass supplies the tighter upper endpoint. Intersecting both ladders leaves nonempty intervals for every family and the exact envelope $c_f \in (0.578, 1.228]$ (Table 2). A preregistered Hill-number audit favors support: leave-one-family-out mean absolute log error is 0.0262 for H_0, versus 0.1839 for H_1, 0.3791 for H_2, and 0.1255 for $|C|k$. Standard transformer arms remain censored; the CoT premium-collapse experiment reaches the $k = 4$ floor on 9/9 chain and 8/9 segmented cell-seeds under the corrected criterion.

Macro compression, allocation shape, and retrieval depth

The uniform-allocation macro control shows why a greedy macro dictionary need not reduce parameters under a uniform per-slot allocation: for `load_mid`, total entries fall 937 → 554 without shrinking the maximum-loaded slot, so constructed free/macro arms coincide in 18/21 cells; the three exceptions are the `hubs3` seeds, where the max-loaded slot shrinks. The slot-elastic construction control changes only the allocation rule; under slot-elastic construction the predeclared prediction error switches from the max-slot model (0.048 versus 0.233 for total-entry compression under uniform allocation) to the total-entry model (0.108

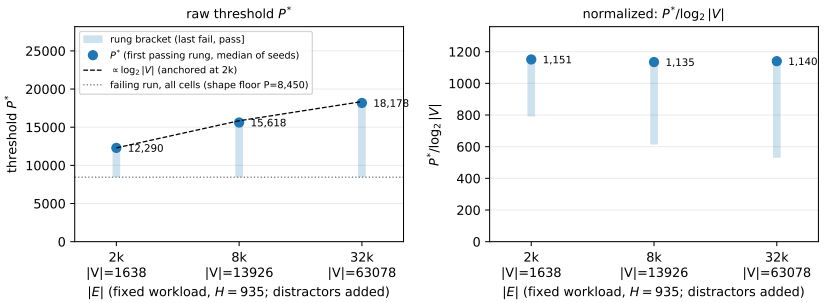


Figure 1: Distractor control for theory-matched capacity scaling. H is held fixed while unused vocabulary grows. All three sizes share the same architectural fail floor $P = 8450$ and first pass at the same coarse rung index; the pass–fail intervals overlap. The control therefore resolves no rung shift but does not identify a $\log_2 |V|$ scaling law.

versus 0.364 under slot-elastic allocation). Thus macro-decomposition credit is a real parameter saving for the explicit construction when budget follows total slot demand.

Training differs. The trained macro-segmentation control lowers the threshold under macro segmentation, but the depth-matched private-macro control holds segmentation depth fixed while making macro identities query-private: private and shared thresholds are exactly equal in 27/30 family-seed-budget cells and within one rung in 29/30; the budget-B family medians have $R_{\text{reuse}} = 1.000$ throughout. On `reuse_hi3`, the private arm stores more associations than the free arm yet reaches criterion with 1.5–2.2× fewer table parameters. The trained gain is therefore attributed to fewer lossy live lookups, not the discrete cross-query reuse credit S_k (Appendix D). Every headline claim, its exact formal scope, and its empirical status are collected in Table 4 (Appendix C).

Main empirical synthesis. Across the uncensored theory-matched experiments, the graph-derived support charge organizes where capacity binds, while the ablations identify which structural term is realized by which mechanism. Support diversity, not frequency concentration, best predicts the threshold ladder; the precise vocabulary-size factor remains unresolved by the distractor control. Macro decomposition saves explicit parameters only when allocation follows total slot demand, and the trained macro benefit survives removal of cross-query reuse, pointing instead to shorter live retrieval depth. The composition experiments add an orthogonal constraint: independently coded cells can each memorize their local closure yet still fail at the crossing because the receiving representation is misaligned. These findings separate resources that raw fact count conflates—stored support, allocation/width, and cross-cell alignment—while keeping architecture-specific scaling factors and open lower-bound extensions explicitly conditional.

5 Conclusion and Future Work

We introduced a graph-parameterized view of parametric memory for multi-hop chain recall, where capacity depends on support usage, decomposition, realizable width, and cross-module alignment rather than raw fact count. The theory gives an explicit upper bound, a restricted lower bound for witness-executing $\mathcal{F}_{\text{pf}}$, and a conditional composition floor; controlled experiments support the structural picture while showing that macro gains depend on allocation and retrieval depth and that independently learned memories can fail at cross-cell composition. The scope remains limited: standard transformer thresholds are censored, width is tested only at $W_{\mathcal{D}} \in \{2, 4, 8\}$, the distractor control does not identify the $\log_2 |V|$ factor, and the receiving-gate mechanism is validated only on controlled exact oracles. Future work should extend the lower bound beyond $\mathcal{F}_{\text{pf}}$ and chain queries, resolve the width and vocabulary-scaling factors, test held-out thresholds, and evaluate the receiving and composition mechanisms on broader models and larger real knowledge graphs; full scope and deferred tests are collected in Appendix E.

Ethics Statement

This work uses synthetic graphs and publicly available datasets, including MetaQA, ComplexWebQuestions, and the de-identified MIMIC-IV demo, solely for controlled knowledge-graph memory experiments. The study does not involve human subjects, and the clinical data are not used for diagnosis, treatment, or clinical decision-making.

Reproducibility Statement

The theorem assumptions and complete proofs are given in the appendices; experimental protocols, amendments, seeds, budgets, and negative results are documented in Appendix E and Appendix J. Reported tables and figures are generated from archived result files by version-controlled analysis scripts; no figure or table is hand-edited.

Generative AI Use Disclosure

Generative AI tools , including Claude Opus 5 and Fable 5,were used throughout this work, and their role is described here in full. They were used to edit and polish the manuscript, including improving wording, clarity, organization, and presentation of both the main text and the mathematical proofs, while preserving the authors' intended technical content and arguments. They were also used to assist with proving mathematical claims and with writing code related to the theoretical and experimental components of the study; all proofs, mathematical statements, and code were subsequently reviewed and verified by the authors. For literature discovery, generative AI was used only to help identify potentially relevant papers and suggest search terms. No generative model produced any data used in this study: the synthetic corpora were generated by deterministic, seeded code with parameters fixed in version-controlled configurations, and the real corpora were public datasets used as released. The authors independently verified the mathematical claims, experimental code and outputs, numerical results, citations, and the final manuscript text, and remain responsible for all content.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws, 2024. URL https://api.semanticscholar.org/CorpusID:269005957.

Mikita Balesni, Tomek Korbak, and Owain Evans. Lessons from studying two-hop latent reasoning, 2025. URL https://arxiv.org/abs/2411.16353.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories, 2024.

Moses Charikar, Eric Lehman, Ding Liu, Rina Panigrahy, Manoj Prabhakaran, Amit Sahai, and Abhi Shelat. The smallest grammar problem, 2005.

Thomas Y Chen. How many parameters for multi-hop? an information-theoretic capacity law for knowledge retrieval in large language models, 2025.

Mingcong Dang, Shengling Geng, Yonghui Xu, Fubo Wang, and Lizhen Cui. Sa-rag: Structured and adaptive retrieval-augmented generation for multi-hop question answering. *Neural Networks*, 204:109201, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2026.109201. URL https://www.sciencedirect.com/science/article/pii/S0893608026006623.

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Atri Rudra, James Y Zou, Azalia Mirhoseini, and Christopher Re. Cartridges: Lightweight and general-purpose long context representations via self-study, 2026.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Reshmi Ghosh, Rahul Seetharaman, Hitesh Wadhwa, Somyaa Aggarwal, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. Quantifying reliance on external information over parametric knowledge during retrieval augmented generation (rag) using mechanistic analysis, 2024. URL https://arxiv.org/abs/2410.00857.

Mohamed Rashad Abdel Hamid, Salsabil Amin El-Regaily, and Mostafa Mahmoud Aref. Can large language models perform retrieval-augmented generation as multi-hop reasoning over knowledge graphs? In *2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 360–365, 2025. doi: 10.1109/MIUCC66482.2025.11196859.

Momchil Hardalov, Gonzalo Iglesias, and Adrià de Gispert. Cartridges at scale: Training modular kv caches over large document collections, 2026.

Yujiao Huang, Ling Yang, Xu-Hua Yang, and Xinli Xu. Retrieval-augmented generation for multi-hop question answering based on structured planning. *ACM Trans. Knowl. Discov. Data*, 20(3), March 2026. ISSN 1556-4681. doi: 10.1145/3789506. URL https://doi.org/10.1145/3789506.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

David Johnston and Nora Belrose. Examining two hop reasoning through information content scaling, 2025.

Michael Kallfelz, Anna Tsvetkova, Tom Pollard, Manlik Kwong, Gigi Lipori, Vojtech Huser, Jeffrey Osborn, Sicheng Hao, and Andrew Williams. MIMIC-IV demo data in the OMOP Common Data Model. *PhysioNet*, June 2021. doi: 10.13026/p1f5-7x35. URL https://doi.org/10.13026/p1f5-7x35. Version 0.9.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers, 2026.

Wen-Sheng Lien, Yu-Kai Chan, Hao-Lung Hsiao, Bo-Kai Ruan, Meng-Fen Chiang, Chien-An Chen, Yi-Ren Yeh, and Hong-Han Shuai. Hyperrag: Reasoning n-ary facts over hypergraphs for retrieval augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 2465–2476. ACM, April 2026. doi: 10.1145/3774904.3792710. URL http://dx.doi.org/10.1145/3774904.3792710.

Jessy Lin, Luke Zettlemoyer, Gargi Ghosh, Wen-Tau Yih, Aram Markosyan, Vincent-Pierre Berges, and Barlas Oğuz. Continual learning via sparse memory finetuning, 2025.

Junhong Lin, Bing Zhang, Song Wang, Ziyan Liu, Dan Gutfreund, Julian Shun, and Yada Zhu. How much reasoning do retrieval-augmented models add beyond llms? a benchmarking framework for multi-hop inference over hybrid knowledge, 2026. URL https://doi.org/10.1145/3770855.3817481.

Xukai Liu, Ye Liu, Jipeng Zhang, Yanghai Zhang, Kai Zhang, and Qi Liu. Layer-order inversion: Rethinking latent multi-hop reasoning in large language models, 2026. URL https://arxiv.org/abs/2601.03542.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, volume 2025, pp. 4166–4207, 2025.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

Martino M. L. Pulici, Cuong Xuan Chu, Evgeny Kharlamov, and Volker Tresp. A storage-retrieval gap in parametric knowledge graph memory, 2026. URL https://arxiv.org/abs/2608.25489.

Shauli Ravfogel, Gilad Yehudai, Joan Bruna, and Alberto Bietti. Geometric factual recall in transformers, 2026.

Clayton Sanford, Bahare Fatemi, Ethan Hall, Anton Tsitsulin, Mehran Kazemi, Jonathan Halcrow, Bryan Perozzi, and Vahab Mirrokni. Understanding transformer reasoning capabilities via graph algorithms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Mingyang Song and Mao Zheng. Model merging in the era of large language models: Methods, applications, and future directions, 2026.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 12–22, 2024.

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1240–1250, 2025.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. Self-routing rag: Binding selective retrieval with knowledge verbalization, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.205. URL https://aclanthology.org/2025.findings-acl.205/.

Zhuoyi Yang, Yurun Song, Kyler G Harris, Iftekhar Ahmed, and Ian Harris. Fine-tuning vs. rag for multi-hop question answering with novel knowledge. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 384–392, 2026.

Yuekun Yao, Yupei Du, Dawei Zhu, Michael Hahn, and Alexander Koller. Language models can learn implicit multi-hop reasoning, but only if they have lots of training data. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9684–9702, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.490. URL https://aclanthology.org/2025.emnlp-main.490/.

Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou. Back attention: Understanding and enhancing multi-hop reasoning in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11257–11272, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.567. URL https://aclanthology.org/2025.emnlp-main.567/.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

A Theory: remarks, examples, and proofs deferred from §3

Remark 1 (Scope and selector optimality of Theorem 2)**.** *The theorem is fundamentally an entry-count lower bound. Part (iii) converts that count to parameters only under the stated explicit-entry representation model; shared weights, superposition, or higher-density encodings that store several logical entries in the same scalar budget lie outside that conversion. Parts (i)–(ii) do not require selector optimality beyond their stated class, coverage, and pair-balancedness hypotheses. Part (iii), which identifies the lower-bound charge with the reference-selector $H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$ used by the upper bound, additionally requires selector optimality and $\beta\varepsilon' \le 1/2$ so that the multiplicative factor remains bounded below by a positive constant.*

That selector-optimality hypothesis is not *established by the main synthetic experiments: the generator gives each synthetic query a unique witness, so selector choice is degenerate and the archived* 1.00 *selector ratios are sanity checks rather than evidence for optimality. The four toy graphs in Table 3 — chain, hub star, shared-macro pair, and parallel-branch diamond — are solved exactly by brute force. On the eight clinical subgraphs, four implemented selector heuristics provide only a finite-class diagnostic; and the k=1 real-KG closures are again degenerate because every witness is a single edge. Thus selector optimality remains an explicit hypothesis, and an exact test on multiplicity-bearing graphs is deferred.*

Definition 7 (Latent crossing interface)**.** *A weight-local Φ has a* latent crossing interface *if, for every retained crossing query, the composed hop probe at the crossing is a fixed function of the latent code M_A emits for the bridge vertex and of B-side structures alone — in particular, the bridge vertex is not re-emitted as tokens and re-encoded through the shared frozen token embedding. Merging, adapter mixing, slot union, and KV concatenation have latent crossing interfaces; a pipeline that re-tokenizes the bridge vertex (CoT-style) is weight-local by Definition 6 but does* not*.*

Remark 2 (What μ does and does not price)**.** *H_k is an order-zero support count. Hubs and long-tailed frequencies enter through retained-closure selection, but repeated use of the same (fact, hop-position) pair does not increase H_k: a pair used once or many times is stored once. A naive mass sum $\sum_{f,i} \Pr[f \text{ is used at } i]$ collapses to expected witness length (and*

to k in the fixed-depth experiments), so it cannot serve as a frequency-sensitive alternative. For diagnostics we instead normalize the retained usage masses

$$u_{f,i} = \sum_{q \in C_\gamma} \mu(q)\mathbf{1}[(f,i) \in w(q)], \qquad p_{f,i} = u_{f,i} / \sum_{f',i'} u_{f',i'},$$

and use the Hill family: ${}^0D = |\{(f,i) : p_{f,i} > 0\}| = H_k$, ${}^1D = \exp(-\sum p_{f,i} \log p_{f,i})$, and ${}^2D = (\sum p_{f,i}^2)^{-1}$. These diagnostics distinguish support diversity from usage concentration without changing the storage bounds, which remain support bounds. We therefore call H_k the coverage-conditioned support usage profile *and reserve "μ-weighted" for the retained-closure selection itself.*

Remark 3. *Hub inflation and macro-decomposition credit are separated by construction: hubs raise $H_k^{\mu,\gamma}$ (one fact used at many positions is charged per position), while a decomposition can lower $H_k^{\mathcal{D}}$ either by collapsing several hops into one macro occurrence or by reusing a macro across queries. The depth-matched private-macro control measures these components separately through $H_k^{\mathrm{priv},\mathcal{D}}$.*

Remark 4 (Greedy–optimal gap)**.** *Since Theorem 1 holds for* every *admissible $\mathcal{D}$, using $\mathcal{D}_{\mathrm{greedy}}$ only loosens the upper bound, while Theorem 2 is stated against $\min_{\mathcal{D}} H_k^{\mathcal{D}}$; the remaining gap is therefore exactly whatever suboptimality the greedy decomposition incurs on this graph- and witness-constrained objective. The problem is structurally reminiscent of smallest-grammar compression, but the one-dimensional string guarantees of Charikar et al. (2005) do not directly transfer, so we make no worst-case approximation-ratio claim for arbitrary graph closures. The experiments report the three-way ablation directly (Appendix J.2, the archived T7 table): the undecomposed charge $H_k^{\mu,\gamma}$ (no macros), the greedy charge $H_k^{\mathcal{D}_{\mathrm{greedy}}}$, and an LZ78 distinct-phrase count (a descriptive reference, not an estimate of the optimum) — greedy sits at or below the LZ78 reference on every measured cell ($H^{\mathrm{greedy}}/\mathrm{LZ} \in [0.64, 1.00]$ on the α–s grid, $[0.50, 0.67]$ on the synthetic-capacity budget-probe families), so greedy compresses at least as far as a sequential phrase count does; the exact minimum is verified by brute force on the toy graphs (Appendix J.2), where greedy attains it.*

The quantities S_{collapse} and S_{reuse} are entry-count properties of the discrete decomposition; by themselves they do not predict parameter savings in gradient-trained soft-gated memories, whose empirical mechanism is tested separately in the depth-matched private-macro control.

The quantities above are concrete enough to compute by hand. Table 3 works them on four ≤ 8-fact toy graphs (the released `metrics` module reproduces these numbers exactly; they double as its acceptance tests): macro decomposition can buy credit by collapse and, where present, cross-query reuse ($S_2 > 0$); hub fan-in alone buys nothing; and parallel branches under a single (head, relation) pair leave H untouched while inflating the width parameter $W_{\mathcal{D}}$ introduced next.

Table 3: Worked toy examples at $k = 2$: the same order of fact budget, different prices. Chain: $0 \to 1 \to 2 \to 3$ with queries $(0, r_0 r_1)$, $(1, r_1 r_2)$. Hub star: three heads into one hub, two spokes out, all six queries. Shared-macro pair: two heads whose 2-hop paths share the final 1-hop tail. Diamond: two heads sharing two parallel (u, m) branches, two suffix relations per branch (all eight queries). Computed by `metrics.py`/`mgeo_construct`.

toy graph	$\lvert E\rvert$	H_2	H_2^{greedy}	S_2	$W_{\mathcal{D}}$
chain	3	4	2	2	1
hub star	5	5	5	0	1
shared-macro pair	3	3	2	1	1
parallel-branch diamond	8	8	8	0	2

Proof sketch (full proof: Appendix F; margin Lemma 2). Draw i.i.d. codes uniform on $\sqrt{d}\,\mathbb{S}^{d-1}$ for entities, macro-ids, and slot tags; retained (macro-)facts sharing a (head, macro,

slot) triple are *merged* into one entry whose value sums the $\leq W_{\mathcal{D}}$ endpoint codes, with key $\kappa = \phi(u)+\phi(m)$ plus a per-block slot offset. The contraction step F.1 rewrites witnesses by $\mathcal{D}$, preserves answers, and reduces entries to $H_k^{\mathcal{D}}$. Lemma 2 (margin): a single union bound over the $O(N_{\mathcal{D}}^2)$ code *pairs* — the code set includes all active slot, relation/macro, value-endpoint, and support-head codes, so hop-1 cross terms are controlled on retained and non-retained queries alike — makes every claim below deterministic. With gate bias $b = 2d/3 + d/(9W_{\mathcal{D}})$ and a saturating-ramp gate, matched probes clear b and every mismatched probe stays below it once $d \geq C_1 W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta)$, with explicit unoptimized constant $C_1 = (36C_0)^2$. Hub entities may key arbitrarily many entries because interference is controlled by the global code count, not per-hub degree. The propagation step F.3 carries the margin across exactly $K_{\mathcal{D}}(q)$ active blocks; sub-threshold margins route to $\perp$. Optimizing $\mathcal{D}$ uses the full $P_{\mathcal{D}}$ expression of Theorem 1; only under workload dominance may it collapse to $O(H_k^{\mathcal{D}} d_{\mathcal{D}})$, and the $H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$ form requires the additional branch-functional specialization. In the unstructured regime (no macro decomposition, uniform witness use), $H_k^{\mathcal{D}} = H_k^{\mu,\gamma} \approx k|E_{\text{used}}|$, and k=2 reproduces the empirical factor-two cost of Johnston & Belrose (2025), while $d = O(\log(\cdot))$ matches the logarithmic-dimension phenomenon of Ravfogel et al. (2026).

Remark 5 (Resource accounting for the macro decomposition)**.** *The construction takes a query's macro sequence as given (Appendix F, model conventions): the map from a relation string to its macro string is fixed by $\mathcal{D}$ at build time and computed outside the real-valued gated memory. The resources are therefore reported separately.* Memory parameters. *Theorem 1 counts $P_{\mathcal{D}}$ real-valued table, code, and shared-map parameters.* Dictionary storage. *If $\mathcal{M}_{\mathcal{D}}$ stores relation strings explicitly, its serialized cost is*

$$B_{\text{dict}}(\mathcal{D}) = O\left(\Big(\sum_{m\in\mathcal{M}_{\mathcal{D}}} |m|\Big)\log|R|\right) \leq O(|\mathcal{M}_{\mathcal{D}}|\, k \log|R|) \text{ bits.}$$

This is dominated by the real-valued table only under an additional encoding-scale condition such as $d = \Omega(k\log|R|)$ together with $|\mathcal{M}_{\mathcal{D}}| = O(H_k^{\mathcal{D}})$; Theorem 1 does not *assume that condition, so dictionary bits are not folded into $P_{\mathcal{D}}$.* Parsing computation. *A deterministic longest-match parse costs $O(k)$ dictionary lookups per query and yields $K_{\mathcal{D}}(q)$ active memory blocks; a relation not covered by a longer macro falls back to a single-hop token. This parse cost is charged separately to inference work (and to c_p in the discussion model of Section B.1). The construction therefore supplies a resource vector $(P_{\mathcal{D}}, B_{\text{dict}}, O(k)$ parse work$)$, rather than claiming that the macro parser or dictionary is free.*

Proof of Proposition 1. (i) is the data-processing ceiling: answers are functions of $\mathcal{G}$, so model–task mutual information is at most the description length of $\mathcal{G}$ (the argument form of Allen-Zhu & Li, 2024 and Chen, 2025). (ii) Under the proposition's position-free functionality assumption, each retained witness's next vertex is the unique successor for its current (u, r) pair. Re-tokenizing this vertex therefore preserves a complete retained witness inductively. The CoT system reduces a query of at most k hops to at most k single-hop lookups against one position-free table of $|E_{\text{used}}|$ entries (Appendix F, F.4). □

Remark 6 (Scope of the matching claim)**.** *The upper/lower "matching" pair (Theorems 1 and 2) is accordingly claimed only on branch-functional closures; on multiplicity-heavy closures the two bounds address slightly different classes, and closing that gap travels with Conjecture 1. Empirically the restriction holds on every main synthetic corpus and on 39 of 40 clinical subgraph runs; one clinical run has the mild exception $W_{\mathcal{D}} = 2$. The clinical graph's maximum out-degree of* 361 *is therefore not itself the table width. The substantive $W_{\mathcal{D}} > 1$ scaling test is the revision-stage forced-width synthetic realization reported in Section 4, at $W \in \{2, 4, 8\}$.*

Definition 8 (β-pair-balanced query distribution)**.** *Fix γ. For a witness selector $w(\cdot)$ and an admissible decomposition $\mathcal{D}$, say a retained labeled (macro-)fact/slot pair is* carried *by a query q of the retained closure if the $\mathcal{D}$-rewriting of $w(q)$ uses it. For $\beta \geq 1$, μ is* β-pair-balanced *if for every such $(w, \mathcal{D})$, every $\tau \in (0, 1]$, and every $Q' \subseteq \mathcal{Q}_k(\mathcal{G})$ with $\mu(Q') \leq \tau$,*

$$\#\{\text{retained labeled (macro-)fact/slot pairs carried by queries in } Q'\} \;\leq\; \beta\,\tau\, H_k^{\mathcal{D}}(\mathcal{C}_k^{\mu}(\gamma), w).$$

Removing τ query mass can then delete at most a $\beta\tau$ fraction of the distinct pairs. A uniform μ over a closure with comparable witness lengths is $O(1)$-pair-balanced; a μ that parks most distinct pairs on queries of negligible total mass is not.

Proof of Theorem 2. (i) Let

$$Q_{\mathrm{ok}} = \{q \in \mathcal{Q}_k(\mathcal{G}) : M(q) \neq \bot,\ M(q) \in \mathrm{ans}(q)\}.$$

By Definition 1,

$$\mu(Q_{\mathrm{ok}}) \geq 1 - \varepsilon - (1-\gamma) = \gamma - \varepsilon = 1 - \varepsilon'.$$

By condition (c) of Definition 5, for every $q \in Q_{\mathrm{ok}}$ the execution resolves a witness $w_M(q)$ ending at $M(q)$, with each slot resolved by its corresponding table. These witnesses define the selector w_M on Q_{ok}.

By (b) the induced map from resolved labeled (macro-)fact/slot pairs to entries is at most one-to-one per slot, and the resolved pairs form the charge $H_k^{\mathcal{D}_M}(Q_{\mathrm{ok}}, w_M)$ of an admissible decomposition $\mathcal{D}_M$. No identification of Q_{ok} with the fixed retained closure, nor of w_M with the reference selector, is made: since $\mu(Q_{\mathrm{ok}}) \geq 1-\varepsilon'$, the charge is at least $\widetilde{H}_k^{\varepsilon'}$ by definition of the triple minimum.

(ii) Fix $w(\cdot)$ and let $(C, \mathcal{D})$ attain the inner minima, and write $C_\gamma = \mathcal{C}_k^\mu(\gamma)$ and $Q' = C_\gamma \setminus C$. The set C need not be contained in the fixed retained closure; only its overlap with C_γ matters. Because $Q' \subseteq \mathcal{Q}_k(\mathcal{G}) \setminus C$,

$$\mu(Q') \leq 1 - \mu(C) \leq \varepsilon'.$$

Apply β-pair-balancedness to this specific Q'. At most $\beta\varepsilon' H_k^{\mathcal{D}}$ retained pairs have all of their carriers in Q'. Every remaining retained pair has a carrier in $C_\gamma \setminus Q' = C_\gamma \cap C \subseteq C$, and is therefore counted in $H_k^{\mathcal{D}}(C, w)$. Hence $H_k^{\mathcal{D}}(C, w) \geq (1 - \beta\varepsilon')\, H_k^{\mathcal{D}}$, and taking minima over $\mathcal{D}$ and then $w(\cdot)$ gives the claim. For (iii), selector optimality identifies the double minimum with $\min_{\mathcal{D}} H_k^{\mathcal{D}}(w_{\mathrm{ref}}) = H_k^{\mu,\gamma} - S_k^{\mu,\gamma}$; the additional condition $\beta\varepsilon' \leq 1/2$ makes the multiplicative factor at least $1/2$, yielding the stated $\Omega((H_k^{\mu,\gamma} - S_k^{\mu,\gamma}) d_{\min})$ parameter bound. In (i), the passage from "answered correctly" to "resolved along some witness" is condition (c) of Definition 5, which supplies the witness and its slot-by-slot resolution; without (c) a correct output need not induce any entry, which is why (c) is part of the class. □

Remark 7 (Pair-balancedness diagnostic and failure mode)**.** *A single-τ diagnostic of the hypothesis was measured on the non-Zipf corpora (greedy kill fraction at $\tau = 0.05$): kill/τ ratios span* 0.31–1.11 *across the uniform-μ synthetic, clinical, and real-KG cells, but reach* 9.86 *on Zipf-μ cells. Values at or below one are the favorable direction for an order-one pair-balance constant. This audit is therefore compatible with $\beta = O(1)$ on the former families at the tested mass scale, but it does not certify the definition's all-τ requirement. The Zipf-μ cells fail even this diagnostic and therefore do not instantiate the matching lower bound; their fully censored threshold results are unaffected. Natural long-tailed query distributions can violate the hypothesis; a counting step robust to mass concentration is open.*

Remark 8 (Mapping Definition 5 to transformers)**.** *Condition (a) is satisfied by architectures whose hop computation is realized by gated key–value retrieval — FFN layers read as key–value memories, and attention value-matrices likewise, per the associative readings underlying Nichani et al. (2025) and Ravfogel et al. (2026); multi-head attention with KV shared across heads or layers still factorizes provided each resolved hop is attributable to some entry. Condition (b) is the substantive restriction: an architecture that arranged a* single *entry to resolve several distinct macro-facts at the same slot — e.g., by placing many facts in one superposed direction and disambiguating downstream — would exit the class, and this is precisely the escape route Conjecture 1 must close or exploit. We know of no construction realizing such super-factorized storage at scale without violating the margin requirements of Lemma 2 in Appendix F, but we do not rule it out; the conjecture's empirical test (the synthetic pretrained-capacity experiment) is designed so that a super-factorized learner would reveal itself as thresholds tracking $|E|$ rather than $H_k^{\mathcal{D}}$.*

Corollary 1 (Coverage-constrained error form)**.** *If the composed system additionally has expected coverage at least γ' over the relative-frame distribution, i.e. $\mathbb{E}_{\mathbf{O}}[\mathrm{abst}(\Phi)] \leq 1 - \gamma'$, then*

$$\mathbb{E}_{\mathbf{O}}[\mathrm{err}_{\mathrm{conf}}(\Phi)] \ \geq\ (1 - p_{\mathrm{ch}}(d) - \delta_{\mathrm{align}}) \cdot \mathrm{cross}_k^{\mathrm{uniq}}(\mathcal{P}) - (1 - \gamma').$$

An error-only lower bound without a coverage constraint is impossible: a system that abstains on exactly the unique-witness crossing queries has zero confabulation on them and pays the entire bound in coverage, which is the faithfulness-versus-coverage dichotomy of Definition 1 operating as designed.

Corollary 2 (The two escapes, derived)**.** *(i)* Re-tokenization*: emitting the bridge vertex as tokens routes it through the shared frozen token embedding — a codebook common to all cells — so CoT composition is not bounded by Theorem 3; this derives Finding 1 vs. CoT of Balesni et al. (2025). (ii)* Co-occurrence*: facts placed in one cell (same training documents) make the crossing internal,* $\text{cross}_k^{\text{uniq}} \to 0$*; this derives their Finding 3. (iii) Joint training with distractor mixing is not weight-local, hence outside the theorem — explaining, rather than contradicting, its success in Hardalov et al. (2026), whose reported near-chance collapse under naive mixing is qualitatively consistent with a large unavoidable-crossing mass. A quantitative post-hoc computation of* $\text{cross}_k^{\text{uniq}}$ *on those setups was pre-registered as experiment the preregistered quantitative crossing follow-up. No the preregistered quantitative crossing follow-up result is reported in this manuscript, so it is treated as a deferred follow-up.*

Remark 9 (Pretraining as an implicit alignment channel (empirical; see §4))**.** *Theorem 3's collapse floor assumes independently coded cell memories with no frame-alignment channel. The receiving-gate alignment audit measures the random-bridge receiving-cap probability near* 0.04 *on the controlled independent-frame oracle. In the branch-functional no-alignment specialization (*$\delta_{\text{align}} = 0$*), this corresponds to a floor near* 0.96*; the archived* $\mathcal{M}_{\text{geo}}$ *slot-union failure* 0.999–1.000 *lies above, and is therefore consistent with, that floor (Table 8). We do not use the legacy LoRA composition rows quantitatively because their current archive render is incomplete. The relevant hypothesis can fail for shared-base adapters: every cell is trained on the* same *pretrained base, whose token embeddings and circuits provide a common representational frame rather than an independent Haar-relative codebook. This is a possible failure of Assumption 1: shared-base alignment need not satisfy the independent-relative-frame receiving-cap model. Re-tokenization instead leaves the latent-crossing-interface class by routing the bridge through shared tokens (Corollary 2(i)). Neither escape is excluded by weight locality alone. Thus successful composition outside these premises does not contradict the conditional floor.*

Assumption 1 (full statement). Fix a retained unique-witness crossing from cell A to cell B. Let $\mathcal{O}_A$ act isometrically on A's private latent coordinates and let $O \sim \mathcal{O}_A$ be Haar-random, independently of B. Under a frame-nonaligning composition (Definition 6), rotating A covaries only the A-inherited structures; the receiving B table, its keys, and any B-frame probe components are not rotated or aligned to O. Condition on the unrotated cell memories and all frame-invariant side information available to Φ. For bridge vertex v, write $x^{(A)}(v) = O\xi_A(v)$. For a correct B-side entry κ define $\text{Cap}(\kappa) = \{x : \langle p(x), \kappa\rangle \geq b\}$ and let Cap^+ be the union over all receiving entries whose values lie on a realizable continuation (at most $W_{\mathcal{D}}$ entries). Assume

$$\Pr_O\Big[x^{(A)}(v) \in \text{Cap}^+ \mid M_A, M_B, \text{frame-invariant side information}\Big] \leq p_{\text{ch}}(d) + \delta_{\text{align}}.$$

Here $p_{\text{ch}}(d)$ is the normalized random-bridge surface measure of the correct-cap union and δ_{align} absorbs any excess alignment. In the branch-functional $W_{\mathcal{D}} = 1$ case, where the relevant bridge state is non-superposed and the correct receiving region is a single cap, a transitive Haar action with no alignment channel has $\delta_{\text{align}} = 0$, and Remark 15 gives $p_{\text{ch}} \leq e^{-\Omega(d)}$. For general superposed bridge states the cap-overlap probability remains the explicit quantity $p_{\text{ch}}(d)$ in the assumption. Rotationally equivariant from-scratch training is a sufficient model for an independent relative frame; a shared pretrained base fixes a common frame and need not satisfy it. The receiving-gate alignment audit measures correct receiving-cap activation directly on receiving-complete independent-frame and shared-frame oracles; the older cross-cell decoder statistic remains only a compatibility diagnostic and is not an estimator of δ_{align}.

Definition 9 (Path-separating gated-lookup memory)**.** *A composed system is a* path-separating gated-lookup memory *if, on every retained query, its hop-i state is the gated*

sum $x_i = \sum_j v_j\, \sigma(\langle p_i, \kappa_j \rangle - b)$ *of Definition 4 over the union of the cells' entries, every later hop is a function of* x_i *alone, and the output head decodes* x_k *against stored value codes or abstains. In addition, at every retained crossing, any state produced solely by entries whose value endpoints are off the selected/realizable witness set is* path-separated*: following the remaining relation suffix from that state cannot reach* $\mathrm{ans}(q)$*. Hence a hop whose gates all return* 0 *forces* $\perp$*, and an off-witness crossing cannot be repaired by later reconvergence. The* $\mathcal{M}_{\mathrm{geo}}$ *slot-union and KV-concatenation constructions used in the modular-composition experiment satisfy this property by their stored-value-code support; generic gated architectures with merging/reconvergence need not. A merged LoRA adapter on a pretrained transformer is not in this class.*

Corollary 3 (Gate-activation bound for general weight-local Φ)**.** *Drop the gated-lookup hypothesis. Under Assumption 1, for every weight-local, frame-nonaligning* Φ *with a latent crossing interface, the expected* μ*-mass (over the relative-frame draw) of retained unique-witness crossing queries on which the correct receiving gate fires at the crossing hop is at most* $(p_{\mathrm{ch}}(d) + \delta_{\mathrm{align}})\mathrm{cross}_k^{\mathrm{uniq}}(\mathcal{P})$*. Whether a failed crossing gate becomes a wrong answer, an abstention, or is repaired downstream depends on the architecture and is not bounded here; the composition experiments of §4 measure the downstream outcome directly on both a path-separating gated-lookup substrate (where Theorem 3 applies conditionally) and a merged-LoRA substrate (which is outside the theorem's class and is reported descriptively).*

B Discussion: the substrate frontier and verifiability

B.1 Discussion: a cost model for the substrate choice

Let

$$C^* = c_1 \min_{\mathcal{D}} \Big[d_{\mathcal{D}}^2 + \big(H_k^{\mathcal{D}} + |V_{\mathrm{heads}}| + |V_{\mathrm{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}\big) d_{\mathcal{D}} \Big],$$
$$d_{\mathcal{D}} = C W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta).$$

be the real-valued parametric sufficiency supplied by Theorem 1. On branch-functional closures whose charge-minimizing decomposition is workload-dominant, this simplifies to the familiar $c_1(H_k^{\mu,\gamma} - S_k^{\mu,\gamma}) \log(N_{\mathcal{D}^*}/\eta)$ form. Let τ be the training cost per parameter, λ the corpus drift rate (fraction of E replaced per unit time, forcing refresh), q the query rate, c_p parametric inference cost per query, c_e the cost per traversal probe, and $\bar{\ell} = \mathbb{E}_\mu[\text{probes to find } w(q)]$ (depth $\times$ effective branching under the traversal policy).

Remark 10 (A cost model and its threshold)**.** *Per-query costs are* $\mathrm{cost}_{\textsf{param}} = \lambda\, \tau\, C^*/q + c_p$ *and* $\mathrm{cost}_{\textsf{trav}} = c_e\, \bar{\ell}$ *(indexing amortized into* c_e*). Hence* **param** *dominates iff*

$$q \;>\; q^* \;=\; \frac{\lambda\, \tau\, C^*}{c_e\, \bar{\ell} - c_p}, \qquad c_e \bar{\ell} > c_p,$$

and **trav** *dominates otherwise (always, if* $c_e \bar{\ell} \le c_p$*). Corollaries: reuse (*$S_k^{\mu,\gamma} \uparrow$*) and low drift lower* q^**, favoring internalization; structural depth and branching raise* $c_e \bar{\ell}$ *but also raise* $H_k^{\mu,\gamma}$ *superlinearly in hub-heavy closures, so deep* sparse *corpora favor traversal while shallow* reused *corpora favor internalization. A hybrid that parameterizes retrieved fragments at query time (Su et al., 2025) pays* $\lambda_f \tau C^*_{\mathrm{frag}}/q_f + c_p$ *per fragment with fragment-level reuse* q_f*, and is optimal exactly when fragment reuse is high while global reuse is low; recurrent-depth computation (Kohli et al., 2026) trades* c_p *upward to reduce* C^**, occupying a third edge of the frontier.* Within this linear per-query cost model *sensitivity is confined to the ratio* $\lambda\tau/c_e$*; all other quantities are measurable from serving logs and the graph.*

The model's austerity is deliberate, and bounds its claims: batching, caching, accelerator utilization, retriever-index maintenance, non-linear training scaling, and drift that itself moves $H_k^{\mu,\gamma}, S_k^{\mu,\gamma}, W_{\mathcal{D}}$ are all outside the model by design, and each can move q^* when admitted. The confinement of sensitivity to $\lambda\tau/c_e$ is a property *of the model*, not of serving systems; the corollary's defensible content is the *ordering* of regimes under matched budgets — which is what the substrate cost-frontier analysis measures — not the scale of q^*.

B.2 Discussion: verifiability across substrates

Definition 10 (Verifier model)**.** *The graph owner publishes either oracle access to $\mathcal{G}$ (edge probes $(u, r) \mapsto \{t : (u, r, t) \in E\}$) or a binding commitment to E (e.g., a Merkle root). A* certificate *for (q, a) is a string w such that a deterministic polynomial-time verifier (or, equivalently for the lower-bound argument below, a zero-error randomized verifier) with the stated access accepts (q, a, w) iff $a \in \mathrm{ans}(q)$, and accepts $(q, \perp, w)$ iff the bounded-depth frontier of q is exhausted; soundness holds against arbitrary provers. We make no bounded-error randomized query lower-bound claim.*

Remark 11 (Certification work is conserved)**.** *Three costs are distinguished:* witness discovery *(finding a path or exhausting a frontier against the edge oracle),* certificate generation *(writing the string w), and* certificate verification *(the verifier's own probes).*

(a) Supplied witness. *If the prover supplies a candidate witness path — a traversal's transcript, or a chain-of-thought that names every intermediate vertex — a sound verifier checks it with k probes (or an $O(k \log |E|)$-size commitment proof) and performs no discovery. Generation is free given discovery; the traversal did the discovery in the course of answering, and a witness-emitting generator did it inside its forward pass.*

(b) Answer only. *If the prover supplies only a, any deterministic sound-and-complete verifier, and any zero-error randomized verifier, must on some instance make $\geq k$ edge probes of its own (Appendix H): for fewer, two graphs agreeing on all probed edges but differing in an unprobed hop edge disagree on* $\mathrm{ans}(q)$ *and the verifier errs on one. This is a worst-case probe lower bound: a latent parametric generator's forward pass makes zero edge-oracle probes and emits no witness, so in the answer-only case the verifier must pay the $\Omega(k)$ edge-access cost on some instance.*

(c) Certified abstention. *Accepting $(q, \perp, w)$ requires evidence that the bounded-depth frontier is exhausted; a traversal's exhausted frontier is that evidence, and an answer-only prover has none, so the verifier must exhaust the frontier itself.*

Thus the verifier model establishes an $\Omega(k)$ worst-case probe requirement for answer-only deterministic/zero-error verification; a supplied witness instead shifts witness discovery to the prover or generator and leaves k verification probes, while certified abstention may require frontier exhaustion. The work is relocated rather than eliminated, but no $\Theta(k)$ bound is claimed for general branching graphs. Commitment schemes can relocate discovery to a proving service; a candidate witness that fails verification returns the verifier to case (b), and parametric abstention without a witness or proving service is calibration-based, with no sound acceptance rule.

Remark 12 (Justifications and succinct proofs)**.** *Two apparent escapes fall inside the observation. (i) A natural-language justification emitted alongside a parametric answer is, in the verifier model, a certificate string w from an untrusted prover; soundness must hold against arbitrary provers, so the adversary argument of Appendix H applies verbatim — unverified prose buys zero probe reduction; it relocates trust, not work. The sharper case is a generator that emits its* witness path *(chain-of-thought naming each intermediate vertex): that string is a candidate certificate, and checking it costs the same k verification probes as a traversal transcript, with no discovery at all. What differs is completeness: when the emitted path is wrong the verifier can only reject, so a sound verifier that never falls back to discovery is incomplete on exactly the generator's confabulated paths, charging coverage; if completeness forces the verifier to recover a missing witness itself, the answer-only adversary bound applies in the worst case. Capacity therefore does not remove discovery work; a witness-emitting generator* performs *it inside the forward pass and hands the verifier a receipt whose soundness, not its cost, is the open question. (ii) Succinct proof systems (SNARK-style) compress* verification *of an already-located witness. In the edge-oracle model, locating a length-k witness from scratch still requires discovering its successive edges; the formal $\Omega(k)$ statement proved here, however, is the answer-only verifier lower bound. Succinct proofs compress the receipt, not that search step. (iii) Amortization shifts* where and how often *the search is paid, never* whether*: ingestion-time proving pays the witness-discovery cost once per witness set and amortizes it over repeat queries, and cached witness paths, batch verification, and precomputed certificate indexes likewise reprice the separation per witness set rather than per query. The observation prices worst-case online discovery and is silent on amortized totals,*

so any systems conclusion drawn from it is a per-query conclusion. Systems that route between parametric answering and retrieval by self-assessed knowledge (Wu et al., 2025) are, in this model, calibration-based abstention: valuable for coverage, but not a sound acceptance rule — exactly the gap Remark 11(c) describes.

C Claims status and notation

This appendix collects the paper's claim-to-evidence map and notation in one place; the tables below are descriptive summaries and do not add assumptions beyond the cited statements.

Table 4: Every headline claim with its exact formal scope and its strongest empirical status.

claim	formal scope	empirical status
Upper bound (Thm. 1) at $H^{\mathcal{D}}$, $d \propto W_{\mathcal{D}}^2$	existence; explicit construction; full count with workload-dominance collapse	entry count ratio is 1.00 on the synthetic families and 0.996 on the clinical KG; (ε, γ) pass at scaled d; under the fixed $C = 256$ comparison trajectories, p is bracketed to $(5/3, 2]$; a free (C, p) law is not identified; constant $\approx 10^3$ loose
Lower bound (Thm. 2)	$\mathcal{F}_{\mathrm{pf}}$ (Def. 5) with witness execution; pair-balanced μ; (iii) under selector optimality, $\beta\varepsilon' \leq 1/2$, and the explicit-entry model	the theory-matched capacity-scaling experiment's separately trained macro-free $\mathcal{F}_{\mathrm{pf}}$ member shows H-linear, budget-stable scaling on 14/14 families; this is empirical consistency rather than a lower-bound proof. Selector optimality is exact only on toy brute force, diagnostic on clinical heuristics, and degenerate on the unique-witness synthetic families
Macro-decomposition credit $H - S$ as an entry saving; parameter savings depend on allocation	total-entry saving in Thm. 1; parameter realization additionally depends on allocation shape; uniform-allocation construction follows max-slot load; the slot-elastic construction crosses over to the total-entry prediction; the depth-matched private-macro control gives private/shared thresholds that are exactly equal in 27/30 family-seed-budget cells (29/30 within one rung), with no resolved reuse benefit in the budget-B family medians	
Positive latent entry-count premium under CoT	Prop. 1	$k = 3$: segmented reaches $r = 1$ under the frozen margin, while full-chain is unresolved under the corrected criterion; $k = 4$: chain reaches $r = 1$ on 9/9 cell-seeds and segmented on 8/9, with one optimization outlier; $k \geq 5$ untested
Composition floor (Thm. 3)	expectation over independent relative frames; unique-witness crossings; frame-nonaligning weight-local Φ, latent crossing interface, path-separating gated-lookup memory, Assumption 1	receiving-complete alignment-audit oracle: independent-frame crossing correct-gate 0.038–0.051 vs random 0.040–0.044 and within 0.82–0.85; shared-frame crossing gate 0.92–0.93. Archived trained union fails its sanity control
Capacity scaling on pretrained substrates	Conj. 1 (open)	bottom-censored through 10^4-edge closures; the fresh-seed MetaQA $k = 1$ onset near 8×10^3 is optimization-limited, while no matched high-H baseline-B50 CWQ-LCC row is available
Out-of-closure behaviour	— (annotation)	generalization tracks reuse (s=0.6: 0.44–0.65), not pair coverage (s=0: ≤ 0.02)

Table 5: Notation.

$\mathcal{G} = (V, E, R)$; $\mathcal{Q}_k$; μ	corpus graph; k-hop chain queries; query distribution
(ε, γ); $\eta = \min(\varepsilon, 1 - \gamma)$	faithfulness / coverage targets; slack
$\mathcal{C}_k^{\mu}(\gamma)$	retained closure: minimum-cardinality query set of μ-mass $\geq \gamma$, with deterministic tie-breaking
H_k $(H_k^{\mu,\gamma})$	support usage profile under fixed w_{ref}: distinct (fact, hop-position) pairs used by retained witnesses
$\mathcal{D}$; $\mathcal{M}_{\mathcal{D}}$; $H_k^{\mathcal{D}}$	macro decomposition; its dictionary; its charge (labeled facts/macro-facts $\times$ slots)
H^{greedy}; $S_k^{\mu,\gamma} = H_k^{\mu,\gamma} - \min_{\mathcal{D}} H_k^{\mathcal{D}}$	greedy charge (reported); macro-decomposition credit
$H_k^{\mathrm{priv},\mathcal{D}}$; $S_{\mathrm{collapse}}, S_{\mathrm{reuse}}$	private-macro charge; collapse/reuse decomposition for fixed $\mathcal{D}$
$K_{\mathcal{D}}(q)$; $K_{\mathcal{D}}$	active macro blocks for query q; maximum effective macro depth
$W_{\mathcal{D}}$	realizable width: max fact-path count from one head through the branch sets
V_{heads}; $V_{\mathrm{values}}(\mathcal{D})$	query-head vertices; retained candidate-output vertices
$\mathcal{F}_{\mathrm{pf}}$; $\tilde{H}_k^{\varepsilon'}$	position-factorized class; its triple-minimized charge
$\mathrm{cross}_k(\mathcal{P})$; $\mathrm{cross}_k^{\mathrm{uniq}}(\mathcal{P})$	selected/reference-witness crossing mass (empirical); unique-witness crossing mass (formal bound)
p_{ch}; δ_{align}	random-bridge correct-cap probability; excess alignment slack (the receiving-gate alignment audit measures receiving-cap activation)
P^*; c_f	first-passing threshold parameter count at (ε, γ); family-specific prefactor constrained by last-fail/first-pass brackets

D Extended results

Synthetic and clinical pretrained-capacity experiments in full

Pretrained-capacity threshold test: censored, not passed

Across all 43 synthetic k-hop-trained grid cells (families $\times$ $k \in \{1, 2, 3, 4\}$, including Zipf-μ and the hubs+reuse product cells; 5 seeds, 215 threshold runs), *every* k-hop-trained arm LoRA threshold sits at the grid floor $r = 1$ with $\varepsilon < 0.05$ at $\gamma = 0.9$ (Table 7). Simultaneously the α-sweep moves H_k^{greedy} by at most 1.75$\times$ (k=4), below the pre-registered 2$\times$ requirement, so the kill criterion for Conjecture 1 returns *not rejected* but is doubly non-binding (Appendix J.1): at 10^3–10^4-fact corpora, a pretrained 0.5B substrate's capacity is never the binding constraint at any tested depth, and the threshold–H relationship is unresolvable at this scale.

The censoring is not mysterious — it is arithmetic the capacity literature already implies. Rank-1 all-module LoRA on Qwen2.5-0.5B trains exactly 549,888 parameters (counted from the archived adapter configuration); at the ≈ 2 bits/parameter storage rate of Allen-Zhu & Li (2024) that is $\approx 1.1 \times 10^6$ bits, while the information content of even the largest pre-registered workload ($H = 1.8 \times 10^3$, $|V| \approx 1.6 \times 10^3$) is $H \log_2 |V| \approx 1.9 \times 10^4$ bits — a factor of $\geq$57 ($\geq$380 at the pilot's $H = 270$). Bottom-censoring at $r = 1$ was therefore foreseeable from the literature this paper builds on: the design underestimated the floor, and the floor sits where capacity theory puts it. The initial real-KG probe below shows the same separation empirically — the same CWQ corpus at the same size floors at $H = 528$ under hub-snowball sampling and lifts to threshold 2 at $H \approx 8.5$–8.9×10^3 under random-edge LCC resampling, at the frozen epoch budget — though the preregistered real-KG replication subsequently identifies such lifts as optimization-limited rather than capacity-limited (Table 11): realized H sets the optimization cost of reaching the criterion at fixed epochs, not a binding capacity.

The trained $\mathcal{M}_{\text{geo}}$ arm from the earlier synthetic and clinical pretrained-capacity experiments exits the position-factorized class through super-factorization (unlike the separately trained macro-free theory-matched capacity-scaling arm) and its grid is dominated by an abstention-calibration plateau (302/355 top-censored), so the $\mathcal{M}_{\text{geo}}$ column reports the Theorem 1 *construction* check instead: entries/$H_k^{\text{greedy}} = 1.00$ (median; per-family means 1.00–1.01) at $d = 512$ on every synthetic family and 0.996 on the real clinical KG. Under the exact margin-Lemma convention (the paper's final constants with the value-restricted $d/2$ decode; the archived grid predates them), the (ε, γ) check passes 310/373 synthetic runs at $d = 512$ — margins are thin on hub-heavy cells at this dimension — and *every* failing run recovers at $d \in \{1024, 2048\}$ (63/63), consistent with the $d \gtrsim W_{\mathcal{D}}^2 \log(H/\eta)$ scaling the theorem requires; clinical passes 40/40 at $d = 512$ with $\varepsilon = 0$. The clinical KG's maximum graph out-degree reaches 361, but this is not the realizable table width: 39 of 40 clinical subgraph runs have $W_{\mathcal{D}} = 1$ and one has $W_{\mathcal{D}} = 2$. Thus natural data only weakly exercises the merged-entry regime. A revision-stage synthetic realization supplies the controlled width-scaling test directly: *wd_parallel* corpora — parallel branches under a single (head, relation) pair with diverging retained suffixes, the Table 3 diamond at scale, shaped so the greedy decomposition cannot fuse the width away — realize table width exactly W at forced $W \in \{2, 4, 8\}$ (5 seeds each). At the W=1-sufficient dimensions $d \in \{256, 512\}$ the construction fails 29 of 30 runs: at $W \in \{4, 8\}$ uniformly *by structural abstention* ($\hat{\gamma} = 0.45$–0.81 with $\varepsilon \leq 0.001$ — collapsing margins route to $\perp$, charging coverage, exactly as the gate design of Lemma 2 intends), and at the borderline W=2 by a mix of coverage shortfalls and near-threshold ε (≤ 0.056; one seed passes outright at $d = 512$). At $d = 256\,W_{\mathcal{D}}^2$ 14/15 cells pass and at $512\,W_{\mathcal{D}}^2$ 15/15 ($\hat{\gamma} = 0.90$, $\varepsilon = 0$). The frozen grid, however, had no rung between 512 and $256\,W_{\mathcal{D}}^2$ at $W \in \{4, 8\}$, so on its own it brackets the exponent only to $p \in (1.5, 2]$ at W=4 and cannot exclude a linear law at W=8. A logged forced-width amendment re-ran the identical corpora and seeds at the intermediate rungs $\{1024, 2048, 4096, 8192\}$ (CPU only; the frozen rungs reproduce the archive's pass/fail pattern exactly on every seed). Table 6 summarizes the resulting forced-width fill: W=4 fails 2048 on 4/5 seeds and W=8 fails 8192 on 5/5 seeds. Under the fixed normalization $d = 256W^p$, these comparison trajectories give

Table 6: Forced-width follow-up. Entries give the number of seeds passing $(\varepsilon, \gamma) = (0.05, 0.9)$ at the decisive dimensions.

$W_{\mathcal{D}}$	lower rung	pass count	next confirmed pass
2	$d = 512$	1/5	$d = 1024$: 4/5; $d = 2048$: 5/5
4	$d = 2048$	1/5	$d = 4096 = 256W^2$: 5/5
8	$d = 8192 = 256W^{5/3}$	0/5	$d = 16384 = 256W^2$: 5/5

$p > 5/3$ from W=8 and $p \leq 2$ from the passes at $256W^2$. Thus the specific $C = 256$ trajectories with $p = 1$ and $p = 1.5$ are inconsistent with the observed grid. These data do not exclude those exponents for other values of C; identifying a free (C, p) pair requires a joint fit over a denser width grid. The lemma's constant is another matter: $d \geq C_1 W_{\mathcal{D}}^2 \log(N/\eta)$ with the paper's unoptimized $C_1 = (36C_0)^2$ predicts 8.9×10^5, 3.7×10^6 and 1.6×10^7 at $W = 2, 4, 8$ against realized onsets of 1024, 4096 and 16384 — a W-independent factor of $\approx 10^3$, all of it in the union-bound constant. At experiment scale the experiment therefore checks consistency with the theorem's width dependence under the chosen normalization rather than estimating an unconstrained power-law exponent.

Table 7: Synthetic and clinical pretrained-capacity experiments: LoRA threshold rank to reach $(\varepsilon, \gamma) = (0.05, 0.9)$, mean over 5 seeds, and the Theorem-1 construction check (entries/H_k^{greedy} at d=512). "1 (floor)" = bottom-censored at the extended grid floor $r = 1$.

Graph family	k=1	k=2	k=3	k=4	construct.
Random (α=∞, s=0)	1 (floor)	1 (floor)	1 (floor)	1 (floor)	1.00
High reuse (s=0.6)	1 (floor)	1 (floor)	1 (floor)	1 (floor)	1.00
Heavy hubs (α=1.5)	1 (floor)	1 (floor)	1 (floor)	1 (floor)	1.00
Hubs + reuse	1 (floor)	1 (floor)	1 (floor)	1 (floor)	1.01
Zipf-μ	1 (floor)	1 (floor)	1 (floor)	1 (floor)	1.00
Clinical subgraphs	–	1 (floor)	1 (floor)	–	0.996

Secondary arms. The one-hop-trained zero-shot arm shows no *reliable* latent composition at any rank: $\varepsilon \approx 0.9$ for the large majority of (rank, seed) combinations. It is not, however, uniformly flat: on the dual dissociation cell, isolated (rank, seed) combinations at $r \geq 128$ reach the criterion outright ($\varepsilon \leq 0.05$ within 250 steps on 3 of the 15 runs at $r \in \{128, 256, 512\}$; post-hoc confirmation grid), non-monotonically in rank and with high seed variance. Zero-shot composition is thus attainable but seed-idiosyncratic — systematic latent composition still does not emerge, the predicted latent premium, but the absolute reading of the premium is too strong. The sporadic emergence is itself weak but real evidence about Conjecture 1's landscape: an outright criterion pass within 250 steps of one-hop training alone means some seeds land where the storage-format escape of Remark 8 is reachable by SGD, if only idiosyncratically. The original CoT evaluation required a revision-stage correction, which we report in full. *As originally scored*, CoT was a step function in k — every k=1 cell learnable at $r = 1$; k=2 mixed (10/55 runs, 0/5 on Zipf-μ); every $k \geq 3$ cell failing in the broader archived diagnostic grid through $r = 512$ (beyond the A3-frozen $r \in \{1, 16, 256\}$ sweep), and still failing at 15× the step budget (one Zipf-μ r=512 run excluded there by the frozen divergence rule rather than scored) — an apparent contradiction of the premium-collapse prediction that survived its own robustness probe. The revision's disentangling ablation (the CoT decode-window diagnostic; amendment A4, Appendix J.1) found the instrument at fault before it found any physics: the frozen decode window (8 new tokens) cannot contain *any* k=3 chain target (measured chain+EOS length 10–16 tokens on every row of both cells), while the archived $k \geq 3$ runs sit at training loss 0.0000 with $\varepsilon \approx 0.9$ — the same contradictory signature that exposed amendment (A1)'s parsing bug: memorized chains, truncated readout, a mid-chain intermediate scored as the answer (the archived exhibit F3). The k=2 "mixed" pattern is the same arithmetic: its 10 passing runs are exactly the two cells whose chains fit 8 generated tokens. Re-measured at k=3 with the window sized to the target and the full rate×format×rank cross (A4; 2 cells × 2 seeds ×

Table 8: Modular composition: archive-verified $\mathcal{M}_{\text{geo}}$ crossing failure mass by cross_k (mean $\pm$ sd over cells/seeds). Each entry is the fraction of retained crossing queries answered incorrectly or abstained. Legacy LoRA composition rows are omitted because the current archive render does not reproduce their previously reported values. These implementation-level outcomes are not a plug-in estimate of Assumption 1's cap-overlap probability.

Substrate	0	0.05	0.1	0.2	0.4
$\mathcal{M}_{\text{geo}}$ slot union	0.00	$1.00 \pm .00$	$1.00 \pm .00$	$1.00 \pm .00$	$1.00 \pm .00$
Joint training (ctrl, $\mathcal{M}_{\text{geo}}$)	0.00	$.14 \pm .09$	$.11 \pm .03$	$.11 \pm .03$	$.10 \pm .01$
CoT escape ($\mathcal{M}_{\text{geo}}$)	0.00	$.11 \pm .07$	$.10 \pm .04$	$.10 \pm .02$	$.10 \pm .01$

12 (r, lr) points per format — 96 runs, no short-circuit; archived figure F1/F7, right): with per-hop segmented supervision the threshold is $r = 1$ (at the grid floor) on *every* cell $\times$ seed, reaching the criterion within 500–2,500 steps (at $r = 1$: 1,000–2,000) — **the premium-collapse prediction holds at $k = 3$ at the grid floor**. The full-chain format memorizes wherever optimization converges (final training loss $\leq 10^{-3}$ at 38/48 grid points, ≤ 0.05 at 43/48; the five chain-format hot-rate runs end at loss 1.47–1.75) yet fails the frozen criterion at every point. On every fully-memorized run at final loss $\leq 10^{-3}$ (38/48) the residual error equals the retained closure's tail mass to four decimal places (0.0997 uniform, 0.0999 Zipf-μ). Five additional runs finish with loss in $(10^{-3}, 0.05]$ and residual error 0.0999–0.1144: the frozen abstention margin $\text{lp}(\text{answer} \mid \text{prompt}) - \text{lp}(\text{unknown} \mid \text{prompt})$ is mis-specified for chain-format output — the model's next tokens after `[A]` are intermediates, so retained queries rank *below* untrained tails and τ abstains the wrong mass. We report those cells as failing the frozen criterion, with that diagnosis; nothing is re-scored post hoc. Learning rate is the one confound that survives as real: at $r = 1$ only 5×10^{-4} passes reliably (4/4 cell-seeds; the A3-frozen 10^{-4} passes 1/4), while 5×10^{-4} at $r = 256$ fails on 4/4 cell-seeds in both formats with terminal loss 1.1–2.3 (spilling to $r = 16$ on one Zipf-μ seed per format) — a non-monotone rank$\times$rate interaction the reduced sweep could not see. A window audit of every other CoT-scored result closes the ledger: k=1 targets fit the frozen window by worst-case arithmetic; k=4 chains categorically cannot (readout-censored, like k=3); the modular-composition experiment's CoT escape chains per-hop one-hop calls, each decoding a single entity in its own window, and is immune by construction; no non-CoT arm is affected (single-span targets, ≤ 6 tokens).

Modular composition experiment in full

The decomposition (Appendix I) shows how the verified $\mathcal{M}_{\text{geo}}$ rows pay the floor: the weight-local slot-union mass is confabulation-dominated (≈ 0.86 confabulation, ≈ 0.14 abstention), whereas the $\mathcal{M}_{\text{geo}}$ joint-training and CoT controls have zero measured crossing confabulation and pay their residual ≈ 0.10–0.14 failure mass in abstention. We do not extend that split claim to the legacy LoRA rows, whose current archive render is incomplete.

The pre-composition cross-cell decoder probes remain descriptive only: median cross decodability is 0.000 against within-cell 0.471 for $\mathcal{M}_{\text{geo}}$, while shared-base LoRA is near parity. The receiving-gate alignment audit replaces that proxy with a direct receiving-support and cap-event audit using two receiving-complete exact controls: independent-frame uses independent codebooks across the two cells, while shared-frame uses the same oracle tables with a shared codebook/frame. independent-frame holds support at 1 and changes only the relative latent frame: crossing correct-gate activation falls to the random-bridge baseline while within-cell activation remains high. shared-frame keeps the same support but restores crossing activation and final correctness. The archived trained union is not a clean mechanism test because its support is incomplete and random bridges already fire nominally correct gates at high rate.

Definition 11 (Cross-cell compatibility diagnostic)**.** *For a composition with bridge-entity set $\mathcal{B}$, the current probe decodes each bridge representation produced by one cell with the other cell's decoder and records the score normalized to the within-cell decode scale. We report the median over $\mathcal{B}$ as a descriptive* cross-cell compatibility diagnostic. *It is **not** an estimator or upper bound for the theoretical δ_{align} in Assumption 1: the theorem's quantity is an excess*

probability of entering the union of correct receiving acceptance caps under the independent-relative-frame model, whereas the diagnostic is a median decoder score. The diagnostic can therefore distinguish the $\mathcal{M}_{\text{geo}}$ and shared-base LoRA representations qualitatively, but it does not validate Assumption 1 or instantiate its probability bound. The receiving-gate alignment audit therefore reports the cap event directly on the exact oracles and retains this decoder score only as a descriptive compatibility measure.

Unconditional mass and crossing decomposition. The archive contains both crossing-conditional and total retained failure. For the $\mathcal{M}_{\text{geo}}$ union, crossing-conditional failure is essentially 1 at every nonzero cross_k, while unconditional error-plus-abstention is 0.70–0.81 (top-cell router) and 0.98–0.99 (sum router), because the latter additionally includes intra-cell error and non-crossing router abstention. For LoRA, crossing-conditional failure is lower, but no theorem-floor overlay is reported from the median cross-decoding score because that score is not an estimator of δ_{align}. The modular-composition experiment therefore separates observed crossing behavior and representational compatibility from the theorem's conditional probability assumption rather than conflating them.

Parametric, traversal, and hybrid cost frontier

At matched budgets the arms order as the cost model predicts under its own accounting: parametric c_p=1 per query with no certificate; traversal uses $\bar{\ell} = 1.79$–3.00 probes per query (mean 2.38 at $c_e = 1$) with certificate validity 1.000; the hybrid decays from 4.5–11 at $N_q = 10^3$ to $\approx c_p$ at 10^6; q^* is ordered by reuse ($0.034 < 0.039$) and rescales exactly with τ. Units are the harness's; q^* has no wall-clock meaning until c_e/c_p and τ are calibrated on a real stack.

Transformer controls for theory-matched capacity scaling

Transformer arms. Both from-scratch stacks are dominated by an architectural minimum-width floor at small H ($P^* \approx 14.6$k at 2 layers, 29–51k at 4), releasing at the largest H toward $P^*/(H \log_2 |V|) \approx 0.9$–$1.7$; their thresholds are budget-dependent at the floor ($B \neq 3B$), and their controls are not $|V|$-flat even with frozen embeddings (wider stacks are needed to separate larger vocabularies at fixed H). No uncensored capacity claim is made for them, and Conjecture 1 stays open. Runs were distributed across A100, H200, L40S and A30 nodes under frozen configurations; the per-task hardware is recorded in the archive manifest. One accounting disclosure: the theory-matched-arm code path is CPU-only, so its scheduler-billed GPU hours were CPU-computed with the accelerator idle; the trained macro-compression grids and all archive re-renders ran on CPU partitions (≈ 950 core-hours).

Coarse/fine prefactor audit. The archived threshold records were re-read at the level of realized parameter counts. The coarse ladder uses targets $H \log_2 |V| \cdot 0.25 \cdot (256^{1/7})^i$ (spacing ≈ 2.208); because $P = 2d^2 + 2kmd + k$ is integer-constrained, the nominal rung multiplier and the realized first-passing ratio need not coincide. A separate fine ladder uses $\times 1.21$ spacing. Each experiment therefore supplies a (last failing, first passing] interval for the family prefactor rather than a point estimate. Table 2 reports medians of those realized intervals and their intersections. None of the seven families has disjoint coarse/fine brackets. The fine interval is narrower for all seven families; for `load_mid` and `hubs3`, the coarse pass nevertheless gives the tighter upper endpoint. The combined envelope is $c_f \in (0.578, 1.228]$. The archived readout is produced by `report/ei_prefactor_audit.py` (commit `ff46858`).

These brackets treat each archived rung-level pass/fail verdict as fixed; the audit does not re-estimate threshold uncertainty from the underlying per-learning-rate runs.

CoT premium collapse at k=4 under a corrected criterion

The k=3 CoT re-measurement above left the full-chain format failing a frozen abstention margin that was mis-specified for chain-format output. The preregistered CoT premium-collapse experiment redefines the chain margin as $\text{lp}(\text{answer} \mid \text{prompt} + \text{generated chain}) -$

Table 9: Macro-compression control. Predicted ratios compare H/H^{greedy} with the architecture's max-loaded-slot ratio. Observed $T^*_{\text{free}}/T^*_{\text{macro}}$ values are shown for the trained macro-compression experiment and the uniform-allocation constructed control. Seeds 0/1/2 are reported at budget B; the constructed arm has no budget axis.

	predicted		observed $T^*_{\text{free}}/T^*_{\text{macro}}$	
family	H/H^{greedy}	max-slot	trained	constructed
`load_mid`	**1.69**	0.97	0.90 / 1.46 / 0.94	1.00 / 1.00 / 1.00
`load_mid3`	**1.40**	0.96	1.50 / 1.50 / 0.88	1.00 / 1.00 / 1.00
`reuse_hi`	1.27	1.00	0.60 / 1.20 / 1.13	1.00 / 1.00 / 1.00
`hubs_reuse`	1.22	0.95	1.06 / 1.06 / 1.00	1.00 / 1.00 / 1.00
`reuse_hi3`	1.21	1.00	1.00 / 1.00 / 1.00	1.00 / 1.00 / 1.00
`hubs3`	1.15	**1.41**	2.17 / 1.37 / 1.33	**1.22 / 1.22 / 1.49**
`load_hi`	1.08	0.94	1.15 / 1.17 / 1.15	1.00 / 1.00 / 1.00

Table 10: Depth-matched private-macro control, table-only threshold ratios (medians over three seeds at budget B). The private-macro arm uses the same macro segmentation as the shared arm but query-private macro identities. Cell-level exact-match counts reported in the text pool both B and $3B$ budgets. Thus $R_{\text{collapse}} = T^*_{\text{free}}/T^*_{\text{private}}$ measures the trained gain from macro segmentation without cross-query reuse, and $R_{\text{reuse}} = T^*_{\text{private}}/T^*_{\text{shared}}$ isolates any additional reuse benefit.

family	R_{collapse} obs.	pred.	R_{reuse} obs.	pred.
load_mid	2.000	1.691	1.000	1.000
load_hi	1.412	1.084	1.000	1.000
load_mid3	1.667	1.328	1.000	1.046
reuse_hi3	2.216	0.942	1.000	1.292
hubs3	1.000	1.014	1.000	1.121

lp(unknown | same), sizes the decode window to the audited maximum chain length (25 tokens against a measured 21), and re-runs both formats at k=4 on three cells ($\alpha = \infty$ uniform; $\alpha = 1.5$ uniform; $\alpha = \infty$ Zipf-μ) $\times$ 3 seeds $\times$ $r \in \{1, 16, 256\}$ $\times$ two learning rates, full grid. The chain format reaches the criterion at r=1 on 9/9 cell-seeds and the segmented format on 8/9 (the hub-heavy cell, seed 1, lifts to r=256 at both rates; reported as an optimization outlier). The k=4 chain result is uniformly at the floor; the segmented format is at the floor on 8/9 cell-seeds, with one documented optimization outlier at r=256. The archived k=3 full-chain rows cannot be re-scored under the conditional margin because neither the required per-query log probabilities nor the adapters were archived. We therefore assign no corrected-criterion pass/fail status to the full-chain k=3 runs. The direct corrected-criterion result is the CoT premium-collapse experiment at k=4; the segmented k=3 result remains reported separately under its frozen criterion.

Held-out closure generalization

The theorems are in-sample by design; the held-out closure generalization experiment characterizes what a memory trained to the retained closure does off it. A seeded 10% hold-out of retained queries, with τ tuned on the in-closure rows only, gives held-out accuracy on answered queries of ≤ 0.02 on every zero-reuse family ($s = 0$; 12 runs, $n_{\text{held}} = 61$–92) and 0.44–0.65 on the reuse families ($s = 0.6$; 6 runs). The clinical subgraphs ($n_{\text{held}} = 4$ per seed) answer 12/12 and 11/12 held-out queries; this is an annotation of the substrate, not a theorem test.

E Extended limitations, reproducibility, and data

Chain queries only (DAG-shaped queries deferred); Theorem 2 covers the position-factorized class (Definition 5), and the general constant-depth case is open (Conjecture 1); Assump-

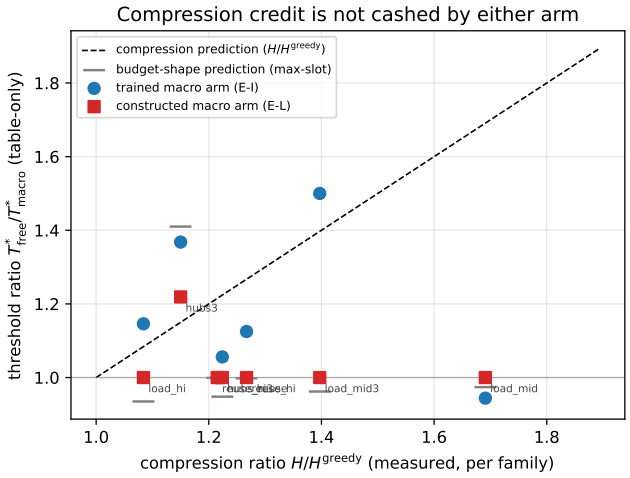


Figure 2: Compression credit vs. realized threshold, all quantities in table-only parameters. The dashed line is the compression prediction H/H^{greedy}; the gray dashes are the architecture's max-loaded-slot ratio. The trained macro arm (circles) and the *constructed* macro arm (squares) both fail to follow the total-entry compression line — including on `load_mid`, where the prediction is 1.69 and the measurement has 2.8 fine rungs of resolution. The constructed control stays at 1 except in `hubs3`, the only family whose max-slot load the dictionary substantially reduces. The credit $S_k^{\mu,\gamma}$ is a saving in *total* table entries; the uniform per-slot budget prices the max-loaded slot, so it generally cannot convert that saving into fewer parameters.

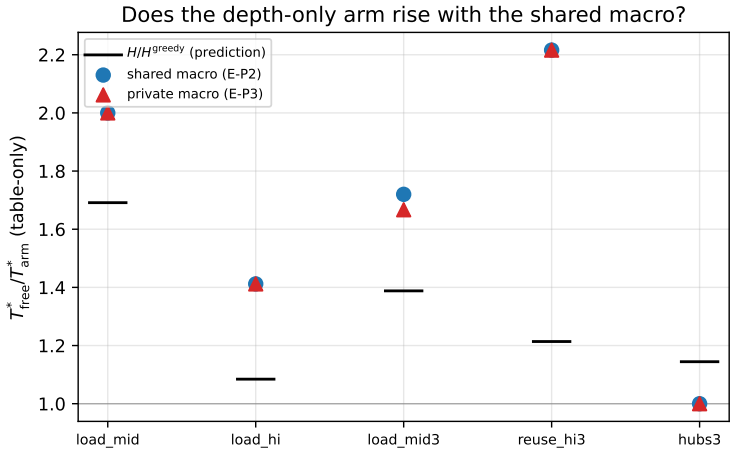


Figure 3: Depth-matched private-macro control: trained shared-macro thresholds (circles) versus the depth-matched private-macro control (triangles), with the total-entry H/H^{greedy} prediction shown per family. The private arm is exactly equal to the shared arm in 27/30 family-seed-budget cells (and within one rung in 29/30); the budget-B family medians have $R_{\text{reuse}} = 1.000$ throughout. In particular, reuse_hi3's private arm stores more associations than the free arm yet reaches criterion with substantially fewer table parameters, so the trained gain is not an S_k association-count saving; together with the private/shared match it points to reduced effective retrieval depth.

tion 1's independent-frame / cap-overlap premise is a model condition; the receiving-gate alignment audit directly measures the receiving-cap event on receiving-complete exact oracles, while the archived trained union fails its own random-bridge and within-cell sanity

checks and is reported only descriptively; Remark 10 inherits the scope of its linear per-query serving-cost model — within that model sensitivity reduces to $\lambda\tau/c_e$, while caching, batching, index maintenance, and non-linear training scaling sit outside it and can move q^*; all capacity claims concern the probed scaling regime of small models; the pre-registered threshold results are bottom-censored (they bound where capacity does NOT bind), and the preregistered real-KG replication strengthens rather than lifts the censoring: the post-hoc MetaQA onset reproduces at the frozen budget but is optimization-limited under the pre-registered budget probe; CWQ-LCC has no matched high-H baseline-B50 1×-step pre-registered real-KG replication row, and the 10^4-edge substrates' query-closure saturation caps the realized H span at 5.7× — testing the top decade requires the 10^5-edge scale, which remains future work. The corrected-readout CoT re-measurement (amendment A4) covers segmented CoT at $k = 3$ under the frozen margin; the CoT premium-collapse experiment gives a uniform floor result for the $k = 4$ chain format and an 8/9 floor result for the segmented format under a corrected, pre-registered chain margin, but the archived $k = 3$ full-chain rows could not be re-scored under it. We therefore assign no corrected-criterion pass/fail status to those $k = 3$ full-chain runs; the two depths are reported under different criteria. The uniform-allocation macro control is scoped to the tested uniform-per-slot $\mathcal{M}_{\text{geo}}$ allocation: the greedy dictionary reduces total entry count but usually not the maximum-loaded slot, so that implementation cannot generally convert $S_k^{\mu,\gamma}$ into a parameter saving. The slot-elastic construction control closes this allocation question for the explicit construction: a slot-elastic budget realizes the total-entry credit. The depth-matched private-macro control shows that the trained macro advantage on the tested families is instead reproduced by a depth-matched private macro arm, with no additional cross-query reuse benefit resolved at the available one-rung resolution. The $W_{\mathcal{D}} > 1$ merged-entry regime is rare in the natural-data checks: every main synthetic corpus and 39 of 40 clinical subgraph runs have $W_{\mathcal{D}} = 1$, while one clinical run has $W_{\mathcal{D}} = 2$. Consequently the controlled width test rests primarily on the revision-stage forced-width realization at $W \in \{2, 4, 8\}$; the matching lower bound's branch-functional specialization applies only to the $W_{\mathcal{D}} = 1$ cases. Every threshold in the paper is measured on the retained closure with the abstention threshold τ tuned to γ on the same rows (§4): the reported (ε, γ) are in-sample by design, appropriate for the storage question Theorem 1 asks but not for a deployment or generalization claim. A separate calibration split and held-out query templates are natural next experiments for the composition and frontier results. Proofs and reported numbers received multiple independent adversarial and archive-backed audits before this version.

Reproducibility and availability

All experiments ran from frozen, version-controlled configurations; every deviation from a frozen protocol is a logged, author-signed amendment (Appendix J.1 and the audit record), and no experiment was ever re-run to move a number — both outcomes of every test were publishable by design. Every result JSON in the archive embeds the SHA-256 hash of its configuration and the git commit that produced it, and the archive is JSON-only and never hand-edited. Tables and figures used in the paper are generated from these records by version-controlled report scripts rather than manually edited; the current figure scripts include `ef_control_figure.py`, `el_figure.py`, and `ep3_readout.py`. The experiments used A100, H200, L40S, and A30 nodes, with the per-task hardware recorded in the archive manifest. The GPU accounting totals ≈ 323 GPU-hours (≈ 292 in the original experimental suite, ≈ 21 in the revision-stage CoT decode-window diagnostic, and ≈ 10 in the preregistered real-KG replication), plus ≈ 60 GPU-hours of guard-verified replay for the archived split table decomposition (re-renders of archived runs, not experiments). Code, frozen configurations, amendment logs, and the complete result archive will be made publicly available upon publication; corpora are regenerated from their public sources by scripts in the release, not redistributed. In this sense the paper's shape is itself the audit trail: every frozen choice, logged amendment, and censored outcome is part of the record it reports.

Data sources. MetaQA (Zhang et al., 2018) is used under CC BY 3.0; the ComplexWebQuestions subgraph union (Talmor & Berant, 2018) derives from Freebase and is used with attribution per its authors' release; the clinical corpus is the openly licensed, de-identified 100-patient `mimic-iv-demo-omop` demo (Kallfelz et al., 2021)(ODbL) — no credentialed

or DUA-restricted data were used in the experiments. All knowledge-graph entities are re-coded to fresh integer surface forms before any model sees them (decontamination by construction, certified by the zero-shot one-hop-transfer probe reported in Appendix D), and no clinical identifier ever enters the repository or the result archive.

F Proof of Theorem 1

Codes. Sample $\phi(x) \sim \mathrm{Unif}(\sqrt{d}\,\mathbb{S}^{d-1})$ independently for every entity, relation, macro-id, and slot tag; normalize so $\|\phi\|^2 = d$. For $x \neq y$, $\langle \phi(x), \phi(y)\rangle/d$ is subgaussian with parameter $O(1/\sqrt{d})$.

F.1 (Contraction). Fix admissible $\mathcal{D}$. Rewrite each retained witness as its $\mathcal{D}$-concatenation; correctness of endpoints is preserved because each labeled macro-fact (u, m, v) is backed by a fact-path from u to v realizing m. The multiset of labeled (macro-)fact/slot pairs has cardinality $H_k^{\mathcal{D}}$ by Definition 3.

F.2′ (Margin)

Codes and the incoherence event. As before, sample $\phi(x) \sim \mathrm{Unif}(\sqrt{d}\,\mathbb{S}^{d-1})$ independently for every entity, macro-id, and slot tag, so $\|\phi(x)\|^2 = d$. Let $V_{\text{heads}} := \{v_0 : (v_0, r_1, \ldots, r_j) \in \mathrm{supp}(\mu)\}$ be the set of query-head vertices in the support of μ. Let Φ be the set of codes the construction ever scores or is scored against: heads and value endpoints of table entries, relation and macro identifiers, the $K_{\mathcal{D}}$ active slot tags, and the head code $\phi(v_0)$ of every $v_0 \in V_{\text{heads}}$. Writing $H := H_k^{\mathcal{D}}$ and

$$N_{\mathcal{D}} := H + |V_{\text{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}} + 1,$$

since $|V_{\text{values}}(\mathcal{D})| \leq H$, we have $|\Phi| = O(N_{\mathcal{D}})$ without any assumption relating $K_{\mathcal{D}}$ to H. The output head scores value codes only (Remark 14), so no dependence on the ambient $|V|$ enters the decode; $|V_{\text{heads}}| \leq |V|$ enters d only through this logarithmic code count.

Lemma 1 (Pairwise incoherence). *Let $N := N_{\mathcal{D}}$, $g := \sqrt{2d\ln(64N^2/\eta)}$ and $E_G := \{\max_{x\neq y\in\Phi} |\langle\phi(x),\phi(y)\rangle| \leq g\}$. Then $\Pr[E_G] \geq 1 - \eta/4$. Moreover $g \leq C_0\sqrt{d\log(N/\eta)}$ with $C_0 = 4$ whenever $\log(N/\eta) \geq 1$.*

Proof. For u uniform on the unit sphere $\mathbb{S}^{d-1}$, $\Pr[|u_1| \geq s] \leq 2e^{-ds^2/2}$; hence for independent codes, conditioning on one of them, $\Pr\big[|\langle\phi(x),\phi(y)\rangle| \geq t\big] \leq 2e^{-t^2/(2d)}$. A union bound over the $\binom{4N}{2} \leq 8N^2$ unordered pairs gives $\Pr[\neg E_G] \leq 16N^2 e^{-g^2/(2d)} \leq \eta/4$ at the stated g. For the absorbed form: $g^2 = 2d\,(2\log(N/\eta) + \ln 64) \leq 2d\,(2 + \ln 64)\log(N/\eta) \leq 16\,d\log(N/\eta)$ once $\log(N/\eta) \geq 1$. □

Every claim below is deterministic on E_G. This replaces any union over (probe, entry) pairs: a false fire is an event over a *pair*, of which there are $\sim H^2$, not H; controlling the $\leq 8H^2$ code pairs once dominates all of them, and — because the bound is now a deterministic statement about the drawn codes — it also covers every *superposed* probe the construction forms at inference time, whose number is query-dependent and need not be polynomially enumerable.

Width of the retained closure.

Definition 12 (Branch sets, realizable frontiers, width). *Fix an admissible $\mathcal{D}$. For head u, macro-or-fact label m, and slot i, let*

$$B_i(u, m) := \{v : (u, m, v) \text{ is used at slot } i \text{ by some retained } \mathcal{D}\text{-rewritten witness}\}.$$

For any *vertex v_0 and* any *macro string $m_1 \cdots m_K$ ($K \leq k$), a* realizable partial witness *of length i is a path $v_0 \to v_1 \to \cdots \to v_i$ with $v_j \in B_j(v_{j-1}, m_j)$ for all $j \leq i$; let $w_i(v_0, m_1 \cdots m_K)$ be the number of such paths and $R_i(v_0, m_1 \cdots m_K)$ the set of their endpoints. Call the pair $(v_0, m_1 \cdots m_K)$* table-realizable through hop i *if $w_i(v_0, m_1 \cdots m_K) > 0$.*

Define

$$W_{\mathcal{D}} := \max_{(v_0,\, m_1 \cdots m_K)} \; \max_{0 \le i \le K} w_i(v_0, m_1 \cdots m_K),$$

the maximum realizable fact-path count over all *table-realizable macro strings — not only those of retained queries. For a retained query q with $\mathcal{D}$-macro string $m_1 \cdots m_{K(q)}$ we abbreviate $w_i(q) := w_i(v_0(q), m_1 \cdots m_{K(q)})$ and likewise $R_i(q)$; the maximum over retained queries is a lower bound on $W_{\mathcal{D}}$. Call the retained closure* branch-functional *if $|B_i(u,m)| \le 1$ for all (u,m,i); then $W_{\mathcal{D}} = 1$.*

Two facts used throughout. First, every realizable partial witness of length $K(q)$ realizes q's relation string (macros are relation strings, and each labeled macro-fact (u,m,v) is backed by a realizing fact-path), so $R_{K(q)}(q) \subseteq \mathrm{ans}(q)$: superposition across genuine same-string branches can never confabulate. Second, q's own selected witness is realizable, so $R_i(q) \neq \emptyset$ at every hop. Multiplicity is irreducible in general — no witness selector removes it (Appendix F) — and $W_{\mathcal{D}}$ is a graph functional computable alongside $H_k^{\mathcal{D}}$.

Model conventions for Definition 4. Blocks are identical up to a per-position offset: block i forms

$$p_i \;=\; \tfrac{1}{\sqrt{3}}\big(\alpha\, x_{i-1} + \phi(m_i)\big) + \beta_i, \qquad \beta_i := \tfrac{1}{\sqrt{3}}\,\phi(i), \qquad \alpha := 1/W_{\mathcal{D}},$$

with α implemented as a fixed scalar multiplier (not the width $W_{\mathcal{D}}$) and the offsets costing $O(kd)$ parameters. The gate is the saturating ramp $\sigma_c(t) = \min(\max(t,0),c)/c$ with $c := d/(9W_{\mathcal{D}})$, i.e. $x_i = \sum_j v_j\, \sigma_c(\langle p_i, \kappa_j\rangle - b)$. The output head scores x_k against the value codes and emits any vertex with score $\ge d/2$ (else $\perp$). As in F.1, the map from a query's relation string to its macro sequence is fixed by $\mathcal{D}$ and computed outside the memory; F.2′ takes the macro sequence as given. The saturating ramp is exactly realizable by two standard ReLU units (Remark 13).

Remark 13 (F.2.5: realization with standard ReLU)**.** *The saturating ramp used above is exactly realizable with two standard ReLU units:*

$$\sigma_c(t) = \frac{1}{c}\left(\mathrm{ReLU}(t) - \mathrm{ReLU}(t-c)\right).$$

Thus saturation is not an additional nonlinear primitive: each gate may be implemented by two ReLU evaluations and fixed affine scaling. This changes only constant factors and introduces no additional asymptotic parameter term.

Construction. One entry per triple (u,m,i) with $B_i(u,m) \neq \emptyset$:

$$\kappa_{(u,m,i)} = \tfrac{1}{\sqrt{3}}\big(\phi(u) + \phi(m) + \phi(i)\big), \qquad v_{(u,m,i)} = \sum_{v \in B_i(u,m)} \phi(v), \qquad b \;=\; \tfrac{2d}{3} + \tfrac{d}{9W_{\mathcal{D}}},$$

placed in T_i for active slots $i \le K_{\mathcal{D}}$. Merging parallel branches only reduces the entry count, so $\sum_i |T_i| \le H$; the value summands total exactly the retained labeled (macro-)fact/slot pairs, i.e. H, and each table value is stored as its summed vector. The table keys and summed values therefore use at most $2Hd$ scalars. Let $V_{\mathrm{values}}(\mathcal{D})$ be the distinct endpoint vertices appearing in those retained values. The output decoder additionally stores one reference code $\phi(v)$ for every $v \in V_{\mathrm{values}}(\mathcal{D})$, costing $|V_{\mathrm{values}}(\mathcal{D})|d \le Hd$ scalars. Learnable head, relation, macro-id, and active-slot codes add $(|V_{\mathrm{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}})d$, and the shared probe/output maps add $O(d^2)$. Thus table keys, summed values, and candidate-output codes together use at most $3Hd$ scalars. A query executes its $K_{\mathcal{D}}(q)$ active slots and halts, so no padding argument and no inequality $k \le H$ is used.

Lemma 2 (Margin)**.** *Assume* $\log(N_{\mathcal{D}}/\eta) \ge 1$*, with $N_{\mathcal{D}}$ as in Lemma 1, and*

$$d \;\ge\; C_1\, W_{\mathcal{D}}^2\, \log\!\big(N_{\mathcal{D}}/\eta\big), \qquad C_1 = (36\, C_0)^2, \qquad (*)$$

with C_0 from Lemma 1 (constants unoptimized). On E_G, for every retained query q and every hop i, simultaneously:

(i) the state satisfies $x_{i-1} = \sum_v m_{i-1}(v)\,\phi(v)$ with support exactly $R_{i-1}(q)$, integer coefficients $m_{i-1}(v) \ge 1$, and $\sum_v m_{i-1}(v) \le w_{i-1}(q) \le W_{\mathcal{D}}$;

(ii) every matched *gate of* T_i *— an entry* (u, m_i, i) *with* $u \in R_{i-1}(q)$ *— saturates,* $\sigma_c(\cdot) = 1$*;*

(iii) every other gate of T_i *is exactly* 0*;*

and the output head emits an element of $\mathrm{ans}(q)$ *— never* $\perp$*, never a wrong vertex. Non-retained queries either abstain or (if their macro string happens to be table-realizable) are answered correctly. Hence, with probability* $\geq 1 - \eta/4$ *over the code draw, the sampled configuration confabulates on* no *query and answers all retained queries; in particular it is* ε*-faithful at coverage* γ*.*

Proof. Condition on E_G throughout. Display $(*)$ is chosen to guarantee the single inequality

$$g \;\leq\; \frac{d}{36\, W_{\mathcal{D}}},$$

since $C_0\sqrt{d\log(N/\eta)} \leq d/(36W_{\mathcal{D}}) \iff d \geq (36\, C_0\, W_{\mathcal{D}})^2 \log(N/\eta)$.

Step 0 (base). $x_0 = \phi(v_0)$: invariant (i) holds with $R_0 = \{v_0\}$ and mass 1.

Step 1 (scores at hop i*, given (i)).* Fix an entry $\kappa = \frac{1}{\sqrt{3}}(\phi(u') + \phi(m') + \phi(i)) \in T_i$. Expanding the inner product over component pairs,

$$\langle p_i, \kappa\rangle \;=\; \frac{1}{3}\Big(\alpha\, m_{i-1}(u')\, d\, \mathbb{1}[u' \in R_{i-1}] \;+\; d\, \mathbb{1}[m' = m_i] \;+\; d\Big) \;+\; X,$$

where the three aligned terms collect the identical-code pairings (the slot pairing $\langle\phi(i), \phi(i)\rangle = d$ is always present, since T_i is scored only by block i), and X collects every pairing of *distinct* codes. On E_G each such pairing is $\leq g$ in absolute value, and the total absolute coefficient mass of distinct-code pairings is at most $\frac{1}{3}\big(3\,\alpha\, w_{i-1} + 6\big) \leq 3$ because $\alpha\, w_{i-1} \leq 1$; hence $|X| \leq 3g$, *uniformly over all entries and all superposed states of mass* $\leq W_{\mathcal{D}}$ — no further union bound is needed.

Matched ($u' \in R_{i-1}$, $m' = m_i$): using $m_{i-1}(u') \geq 1$,

$$\langle p_i, \kappa\rangle \;\geq\; \frac{2d}{3} + \frac{d}{3W_{\mathcal{D}}} - 3g, \qquad \text{so} \qquad \langle p_i, \kappa\rangle - b \;\geq\; \frac{2d}{9W_{\mathcal{D}}} - 3g \;\geq\; \frac{5d}{36W_{\mathcal{D}}} \;\geq\; c,$$

and the gate saturates, contributing exactly $v_{(u', m_i, i)}$.

Unmatched. Distinct entries of T_i carry distinct pairs (u', m') by construction (one entry per triple), so exactly three cases remain: $m' = m_i$, $u' \notin R_{i-1}$ (aligned mass $2d/3$); $m' \neq m_i$, $u' \in R_{i-1}$ (aligned mass $\frac{1}{3}(\alpha\, m_{i-1}(u') + 1)d \leq 2d/3$, since $\alpha\, m_{i-1}(u') \leq \alpha\, w_{i-1} \leq 1$); and $m' \neq m_i$, $u' \notin R_{i-1}$ (aligned mass $d/3$). In every case

$$\langle p_i, \kappa\rangle \;\leq\; \frac{2d}{3} + 3g \;\leq\; b - \frac{d}{36W_{\mathcal{D}}} \;<\; b,$$

and the gate is exactly 0. (Merging parallel branches is essential: without it, two retained parallel branches (u, m, v_1) and (u, m, v_2) at one slot would produce entries with identical keys and invalidate the unmatched-gate bound. The dangerous quantity in the second case is the *coefficient* $m_{i-1}(u')$, controlled by the probe scaling α; without that scaling, a frontier vertex of multiplicity ≥ 2 with an off-macro entry can fire spuriously.)

Step 2 (induction). By (ii)–(iii), $x_i = \sum_{u \in R_{i-1}} v_{(u, m_i, i)} = \sum_v m_i(v)\, \phi(v)$ with $m_i(v) = \#\{u \in R_{i-1} : v \in B_i(u, m_i)\} \geq 1$ exactly on $R_i(q)$. Each such pair (u, v) extends a realizable partial witness ending at u, and distinct pairs yield partial witnesses of length i differing in their last edge, so $\sum_v m_i(v) \leq w_i(q) \leq W_{\mathcal{D}}$: invariant (i) holds at hop i.

Step 3 (decode). For $v^* \in R_{K(q)}(q)$: $\langle x_{K(q)}, \phi(v^*)\rangle \geq d - W_{\mathcal{D}}\, g \geq \frac{35d}{36} > d/2$. For a value code $y \notin R_{K(q)}(q)$: $|\langle x_{K(q)}, \phi(y)\rangle| \leq W_{\mathcal{D}}\, g \leq \frac{d}{36} < d/2$. Emitted vertices therefore lie in $R_{K(q)}(q) \subseteq \mathrm{ans}(q)$, and $R_{K(q)}(q) \neq \emptyset$ rules out $\perp$ on retained queries.

Step 4 (non-retained queries; missing entries). If at some hop no entry of T_i is matched, every gate is 0 by Step 1, so $x_i = 0$; then all subsequent probes have aligned mass $\leq 2d/3 <$

$b - 3g$, all later states are 0, the head sees maximal score $0 < d/2$, and the system emits $\perp$ — charging coverage, never faithfulness; by Definition 2 the abstained mass is $\leq 1 - \gamma$. If instead the pair $(v_0, m_1 \cdots m_K)$ is table-realizable throughout, Steps 0–3 apply with q replaced by $(v_0, m_1 \cdots m_K)$, for two reasons that deserve record. First, the hop-1 cross terms are controlled on E_G because $\phi(v_0) \in \Phi$ for every $v_0 \in V_{\text{heads}}$ (the extended code set of Lemma 1). Second, the width invariant $\alpha\, w_{i-1} \leq 1$ consumed by Step 1 holds *off* the retained closure because Definition 12 maximizes $W_{\mathcal{D}}$ over all table-realizable strings: branch sets are shared, and a non-retained string that mixes prefixes of several retained witnesses is itself table-realizable, hence an instance of that maximum — so $\sum_v m_{i-1}(v) \leq w_{i-1}(v_0, m_1 \cdots m_K) \leq W_{\mathcal{D}}$ propagates by the Step-2 induction verbatim. The answer is then correct: any realizable path realizes the query's own relation string, so $R_K(v_0, m_1 \cdots m_K) \subseteq \text{ans}(q)$.

Step 5 (probability accounting; interface to F.3). Everything above is deterministic on E_G, and $\Pr[E_G] \geq 1 - \eta/4$. No additional propagation failure event is needed: on E_G every non-fired gate contributes exactly 0. Thus the only probabilistic failure event used by the construction is E_G^c, whose probability is at most $\eta/4$; F.3 reduces to the $\perp$-routing above plus the full parameter count $2H_k^{\mathcal{D}} d + |V_{\text{values}}(\mathcal{D})| d + (|V_{\text{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}) d + O(d^2)$, with $|V_{\text{values}}(\mathcal{D})| \leq H_k^{\mathcal{D}}$. F.4 (the CoT reduction) is unaffected except that it inherits the same bias and gate conventions. □

Restatement of Theorem 1. For every admissible decomposition $\mathcal{D}$, let $K_{\mathcal{D}} = \max_q K_{\mathcal{D}}(q)$, $N_{\mathcal{D}} = H_k^{\mathcal{D}} + |V_{\text{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}} + 1$, and $d_{\mathcal{D}} = C W_{\mathcal{D}}^2 \log(N_{\mathcal{D}}/\eta)$. Since $|V_{\text{values}}(\mathcal{D})| \leq H_k^{\mathcal{D}}$, this $N_{\mathcal{D}}$ also dominates the candidate-output code count. The construction above is ε-faithful at coverage γ and uses

$$P_{\mathcal{D}} = O\big(d_{\mathcal{D}}^2 + (H_k^{\mathcal{D}} + |V_{\text{heads}}| + |V_{\text{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}) d_{\mathcal{D}}\big)$$

real-valued parameters. Optimizing $\mathcal{D}$ therefore means minimizing this *full* expression jointly over charge, width, effective macro depth, and codebook size. If $\mathcal{D}$ satisfies workload dominance $H_k^{\mathcal{D}} \geq d_{\mathcal{D}}$ and $H_k^{\mathcal{D}} \geq |V_{\text{heads}}| + |V_{\text{values}}(\mathcal{D})| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}}$, then and only then does the bound simplify to $O(H_k^{\mathcal{D}} d_{\mathcal{D}})$. If, in addition, $W_{\mathcal{D}} = 1$ and a charge-minimizing $\mathcal{D}^*$ satisfies those conditions, the familiar branch-functional form $O((H_k - S_k) \log(N_{\mathcal{D}^*}/\eta))$ follows. Dictionary bits and parser work are the separate resources of Remark 5; no inequality $d \geq k \log |R|$ is implied by the dimension bound.

Remark 14 (Decode over value codes)**.** *The output head stores and scores only the $|V_{\text{values}}(\mathcal{D})| \leq H$ candidate endpoint codes. Vertices outside the retained range are never emitted — consistent with structural abstention (charging coverage, not faithfulness) — and no $\log |V|$ enters d.*

Remark 15 (Receiving-cap margin)**.** *For the crossing-gate argument in Appendix G, on E_G a cross-cell probe whose only aligned components are macro and slot fires the correct B-side gate only if*

$$\tfrac{1}{3}\,\alpha\,\langle x, \phi(u)\rangle \;\geq\; b - \tfrac{2d}{3} - 3g \;\geq\; \tfrac{d}{36 W_{\mathcal{D}}}, \qquad \textit{i.e.} \qquad \langle x, \phi(u)\rangle \;\geq\; \tfrac{d}{12}.$$

For a non-superposed bridge state, for example in the branch-functional $W_{\mathcal{D}} = 1$ case, $\|x\| = \sqrt{d}$ and this gives the constant angular margin $\cos\theta \geq 1/12$; a single receiving cap therefore has Haar measure $e^{-\Omega(d)}$ by standard spherical concentration (Vershynin, 2018).. For a general superposed state with $\|x\| = O(W_{\mathcal{D}}\sqrt{d})$, the same inner-product requirement gives only $\cos\theta = \Omega(1/W_{\mathcal{D}})$, so the corresponding cap measure scales as $e^{-\Omega(d/W_{\mathcal{D}}^2)}$ (with a union factor for multiple receiving caps). Theorem 3 uses the cap-overlap probability as an explicit assumption, so this broader superposed-state observation is not needed for its unique-witness statement.

F.3 (Propagation and abstention). On E_G non-fired gates contribute exactly 0 (Lemma 2, Steps 1 and 5), so a matched block outputs the exact merged value: there is no residue term and no separate probability budget. If at any block no gate reaches half-margin, the head outputs $\perp$; by Definition 2, only non-retained (μ-tail) queries lack entries, so abstentions have mass $\leq 1 - \gamma$ and confabulation mass is $\leq \varepsilon$. Parameter count:

$2Hd + |V_{\text{values}}(\mathcal{D})|d + (|V_{\text{heads}}| + |R| + |\mathcal{M}_{\mathcal{D}}| + K_{\mathcal{D}})d + O(d^2)$, with $|V_{\text{values}}(\mathcal{D})| \le H$, which is exactly the $P_{\mathcal{D}}$ accounting in Theorem 1. Taking the minimum of the full $P_{\mathcal{D}}$ expression over $\mathcal{D}$ yields the theorem; taking only the minimizer of $H_k^{\mathcal{D}}$ is justified only in the stated branch-functional workload-dominant specialization.

F.4 (CoT reduction, for Proposition 1). Assume the position-free used-fact table is functional: $|\{v : (u, r, v) \in E_{\text{used}}\}| \le 1$ for every (u, r) (equivalently, $W_1 = 1$ for a nonempty table). This condition concerns the union across all hop positions and is stronger than functionality within each separate slot. With intermediate emission, each hop's unique output vertex is re-tokenized. Use one position-free table of all used facts with a single fixed dummy tag ϕ_0 shared by every lookup:

$$\kappa_{(u,r)} = \frac{1}{\sqrt{3}}\big(\phi(u) + \phi(r) + \phi_0\big), \qquad p(u, r) = \frac{1}{\sqrt{3}}\big(\phi(u) + \phi(r) + \phi_0\big).$$

The same table is reused at every hop; ϕ_0 is not a position tag and costs only one additional d-dimensional code. A matched lookup has aligned score d (up to the controlled cross-terms), whereas a mismatched key has at most two aligned components and therefore aligned score at most $2d/3$. Hence the $W = 1$ gate convention

$$b = \frac{2d}{3} + \frac{d}{9}, \qquad c = \frac{d}{9}$$

separates matched and mismatched gates by the same margin as in Lemma 2. Each table value contains the single endpoint code of its unique successor, so decoding uses $d = O(\log |E|)$ by the same incoherence argument. For every retained query, induction shows that each decoded successor is the next vertex of its retained witness; hence every remaining lookup exists. For a non-retained query, completing all lookups yields a genuine witness, while a missing lookup causes abstention. Re-tokenization lets every hop reuse the same table, giving $O(|E_{\text{used}}| \log |E|)$ storage up to lower-order code terms. If $W_1 > 1$, decoding one valid successor need not preserve a completion for the remaining suffix, so this argument does not establish the same guarantee without an additional branch-selection or search procedure and its resource accounting.

G Proof of Theorem 3 and Corollary 3

G.1 (Crossing-gate activation). Fix a weight-local, frame-nonaligning Φ and a retained unique-witness crossing query q whose unique witness moves from cell A to cell B at bridge v at hop i. By the latent interface hypothesis, the crossing probe is a fixed function of the emitted A-side bridge code and B-side structures. Under the componentwise covariance in Definition 6, replacing A by OM_A rotates the A-inherited bridge representation to $x^{(A)}(v) = O\xi_A(v)$ but does not rotate the B receiving keys or estimate/undo O. Conditional on the unrotated memories and frame-invariant side information, the correct receiving region Cap^+ is therefore fixed in B's frame while the bridge code carries the independent Haar relative frame. Assumption 1 gives

$$\Pr_O[\text{correct receiving gate fires} \mid M_A, M_B, \text{frame-invariant side information}] \le p_{\text{ch}}(d) + \delta_{\text{align}}.$$

Integrating over the conditioned quantities and summing retained unique-witness crossing μ-mass gives Corollary 3. This is the only use of the frame assumption. An operator that estimates a relative frame and aligns the cells is outside the frame-nonaligning class rather than being silently absorbed into an invariance claim.

G.2 (Absorption in a gated-lookup memory). Now let the composed system be a gated-lookup memory (Definition 9) and consider the crossing hop of this unique-witness query q. Exactly one of three things happens. (a) The correct entry fires: probability $\le p_{\text{ch}} + \delta_{\text{align}}$ by G.1. (b) No entry fires: the state is zeroed, every later probe has aligned mass below b (Lemma 2, Step 4), and the head emits $\perp$ — the query is abstained. (c) Only entries whose value endpoints lie off every realizable partial witness of q fire: the state has no mass on $R_i(q)$; by the Definition 9 hypothesis every later hop is a function of that state, so by induction x_k has no mass on $R_k(q) \subseteq \text{ans}(q)$; the head scores only stored value codes

(Remark 14), so it either abstains or emits a vertex outside ans(q) — the query is abstained or confabulated. Because q has a unique realizable witness, there is no alternative within-cell witness that can bypass this crossing. Hence, with probability taken over the relative-frame draw, for each retained unique-witness crossing query $\Pr[\text{confabulate}] + \Pr[\perp] \geq 1 - p_{\text{ch}} - \delta_{\text{align}}$; summing μ-mass over those queries gives the expected bound in the theorem. The coverage-constrained form is Corollary 1.

H The probe lower bound of Remark 11(b)

The argument below is for deterministic verifiers; fixing the random coins gives the same conclusion for zero-error randomized verifiers. We make no bounded-error randomized lower-bound claim here.

Adversary argument. Suppose verifier V is sound and complete and on instance (q, a, w) makes $t < k$ edge probes. The probes touch at most t of the k hop positions of every witness of q; pick an unprobed position i^* and construct $\mathcal{G}'$ from $\mathcal{G}$ by rerouting the i^*-th edge of the (unique, by construction of the hard instance) witness to a fresh vertex, leaving all probed answers unchanged. $\text{ans}_{\mathcal{G}}(q) \neq \text{ans}_{\mathcal{G}'}(q)$, yet V's view is identical on $(\mathcal{G}, \mathcal{G}')$, so V accepts the same (q, a, w) on both and is unsound on one. Hence $t \geq k$ on some instance. For commitment-based proofs, the k Merkle openings substitute for probes, but locating which k edges to open is witness discovery — a traversal — performed by the prover; the work is relocated, not removed.

I Extended results

The preregistered real-KG replication sweep and the composition split are reported here; the post-hoc initial real-KG capacity probe tables are in the archive.

Preregistered real-KG replication: the onset is optimization-limited, and censoring strengthens

The confirmatory successor was pre-registered in full (frozen protocol, config, outcomes, and dual epoch-matched budgets; amendment record round 4) and executed on fresh cells: seeds $\{10, 11, 12\}$ disjoint from every prior initial-probe seed, so substrates, query samples, and training are all new measurements. H was scaled by query load on the same two real decontaminated substrates, with a pre-registered budget probe: every rung at B50 (50 epochs, the frozen sizing) and the upper rungs additionally at B150.

Table 11: Preregistered real-KG replication: k-hop-trained arm threshold per seed at epoch-matched budgets. One design limitation surfaced and is disclosed: on the 10^4-edge substrates the k=1 query closure saturates near $H = 8.1 \times 10^3$ on MetaQA and $H = 8.6$–8.8×10^3 on CWQ-LCC, so the highest planned query loads no longer expand the realized support as intended (overall span 5.7×, short of the planned ≥10×; the top decade requires the 10^5-edge substrate scale). The saturated B50 rung consequently runs 2× the steps at the same H — an unplanned intermediate-budget replicate.

Substrate	n_q	realized H	budget	threshold (s10/s11/s12)
MetaQA	1 600	1 440	B50	1 / 1 / 1 (floor)
MetaQA	4 000	3 600	B50	1 / 1 / 1 (floor)
MetaQA	10 000	8 070–8 195	B50	**2/2/2**
MetaQA	20 000 (sat.)	8 070–8 195	B50 (2× steps)	1 / 1 / 1 (floor)
MetaQA	10 000	8 070–8 195	B150	1 / 1 / 1 (floor)
MetaQA	20 000 (sat.)	8 070–8 195	B150	1 / 1 / 1 (floor)
CWQ-LCC	4 000	3 600	B50	1 / 1 / 1 (floor)
CWQ-LCC	20 000 (sat.)	8 614–8 793	B50 (2× steps)	1 / 1 / 1 (floor)
CWQ-LCC	20 000 (sat.)	8 614–8 793	B150	1 / 1 / 1 (floor)

On MetaQA, the baseline-B50 onset reproduces on fresh seeds (threshold 2 on 3/3 at $H \approx 8.1 \times 10^3$) and returns to rank 1 on 3/3 at B150. CWQ-LCC has no high-H baseline-B50 1×-step row in preregistered real-KG replication; its saturated high-H B50 row uses 2× steps and is already at rank 1. Across the tested single-hop closures, adequate training budget therefore removes the observed threshold lift. These experiments do not test multi-hop latent recall on the real KGs.

Confabulation/abstention split (archived supplementary split table). The $\mathcal{M}_{\text{geo}}$ slot-union crossing failure is ≈ 0.86 confabulation and 0.14 abstention at every cross_k (guard-verified replay, bit-exact on 58/75 joint-control points); the $\mathcal{M}_{\text{geo}}$ joint-training and CoT controls show zero measured crossing confabulation and pay their residual mass in abstention. LoRA splits are not published because GPU replays were not bit-reproducible at near-threshold scores, so no zero-confabulation claim is made for those rows.

J Experimental amendments and secondary-arm analyses

J.1 Logged amendments (summary)

A1: EOS placed inside supervised targets after a smoke run scored hallucinated continuations (pre-fix artifact preserved). A2: rank grid extended to $r = 1$ before freeze. A3: one-hop-transfer/CoT sweep reduced to $r \in \{1, 16, 256\}$ at the dominant rate. A4: the frozen 8-token CoT decode window cannot contain any $k \geq 3$ chain (measured 10–16 tokens); archived $k \geq 3$ CoT rows are readout-censored and were re-measured with the window sized to the target (opt-in keys; frozen behaviour reproduced byte-identically). A7.1: forced-width d-grid fill at intermediate rungs. A7.2: trained macro-compression arm, frozen before launch. Full designs and amendment text are in the archive; the pre-registered synthetic-capacity kill criterion: if across the α-sweep at matched $|E|$ the threshold varies by less than seed dispersion while H_k^{greedy} varies by $\geq 2\times$, Conjecture 1 is rejected.

J.2 Compression ablation: no macros / greedy / LZ78, and selector spread

Archive-count bookkeeping. The denominators used in the appendix refer to different archived analyses: 355 trained-arm threshold runs, 373 explicit-construction (ε, γ) evaluations, and 353 synthetic cells entering the selector diagnostic. They are not interchangeable run counts.

The compression audit is associated with the 355-run $\mathcal{M}_{\text{geo}}$ archive. The current `report.build` export `t7.csv` is only a one-row stub and is therefore not described as the full per-cell table here. The archived analysis compares the undecomposed charge H (no macros), the greedy charge H^{greedy} (the quantity every experiment reports), and the LZ78 distinct-phrase count over retained witness streams — a *descriptive* sequential-compression statistic with no formal relation to the graph-constrained minimum $\min_{\mathcal{D}} H_k^{\mathcal{D}}$, reported as a reference point and not as an estimate of that optimum. The archived analysis summary reports that greedy compresses (H^{greedy}/H down to 0.65 at $k = 4$; exactly 1 at $k = 1$, where no macro exists to share) and is no less compressive than the descriptive LZ78 reference on the reported families ($H^{\text{greedy}}/\text{LZ} \in [0.64, 1.00]$ on the α–s grid and $[0.50, 0.67]$ on the budget-probe families). Because the current `t7.csv` export is a stub, these ranges are reported as archived analysis summaries rather than as values independently reconstructed from that export. Only the toy graphs certify exact optimality; these LZ78 comparisons do not estimate the graph-constrained greedy–optimal gap. The last column reports the archived per-cell selector diagnostic over four witness selectors (greedy-H, random, lexicographic-shortest, crossing-minimizing). On the 353 synthetic cells the value is exactly 1.000, but this is *vacuous evidence*: the synthetic generator enforces a unique witness for each query, so every selector must coincide. The four toy graphs in Table 3 — chain, hub star, shared-macro pair, and parallel-branch diamond — are the only current exact brute-force selector check. Let $\mathcal{W}_4$ denote the four tested selectors. On the frozen clinical subgraphs, $\min_{w \in \mathcal{W}_4} H_k(w)/H_k(w_{\text{ref}}) = 1.000$, with the tested alternatives strictly worse (random 1.63–2.76×; shortest-lexicographic and crossing-minimizing 1.05–

1.41×). This is a useful finite-class diagnostic, not the minimum over all valid selectors. On the preregistered real-KG closures (k=1) all four again coincide by the single-edge degeneracy, and the recomputed H reproduces the archived real-KG replication ranges (MetaQA 8070–8195; CWQ-LCC 8614–8793). (An earlier rendering pooled distinct cells before taking the minimum, producing spurious large spreads; the per-cell computation here replaces it.)