How Should Reasoning Be Organized in a Transformer's Latent Space?

Anonymous authors
Paper under double-blind review

Abstract

Continuous reasoning has emerged as a promising way to improve reasoning in large language models (LLMs). Yet we still lack a clear principle for deciding what a latent state should preserve. Reasoning by superposition shows that a single latent state can encode several search alternatives and expand them in parallel. We ask how those states should be weighted as reasoning proceeds. A natural choice is to preserve only the states active at the frontier step, since keeping every reached state appears to spread a limited hidden width too thin. We show that the opposite can hold. When later computation draws on several reached states, a cumulative state can guide attention correctly at a smaller hidden width than a frontier state that stores fewer states. At the same width, the cumulative state therefore keeps more intermediate states available for later reasoning. More generally, equal cumulative weights are optimal when future queries are unknown and remain close to the best task-specific weights when those queries are known. Experiments with two-layer and GPT-2 Transformers reproduce the predicted width advantage and show that unequal weights fail first on the states that receive the least weight. This suggests a important principle: keep reached states equally weighted, and restore equal weights as computation proceeds.

1 Introduction

What should a latent state preserve as reasoning unfolds? This question matters as language models move part of their reasoning from written tokens into continuous computation. Continuous chain-of-thought feeds latent states back into the model (Hao et al., 2024); recurrent Transformers update a latent state over several steps (Xu & Sato, 2024; Fu et al., 2026); and reasoning by superposition shows that one latent state can represent several search alternatives at once (Zhu et al., 2025b). These results establish that superposed reasoning is possible, but they leave open what the latent state should preserve as computation continues.

We compare three choices. A *frontier state* keeps only the states active at the current step. A *cumulative state* keeps all states reached so far with equal weight. A general *weighted state* keeps the same set of reached states but assigns them different nonnegative weights. All three use the same task, candidate answers, Transformer, and final query.

The usual capacity intuition favors the frontier state. When more states share one normalized latent state, each receives less weight and should become harder to recover (Elhage et al., 2022). Reasoning, however, rarely uses one earlier state in isolation. A later step may combine information from several reached states. Their useful components reinforce the correct answer, while unrelated candidates remain random interference. This leads to our central result: although it stores more states, the cumulative state can work at a smaller hidden width than the frontier state.

Cumulative-state hypothesis. When any reached state may be needed again, the latent state should assign equal weight to all reached states. Let $\mathfrak{D}(\boldsymbol{a};\mathcal{U})$ denote the smallest hidden width that meets the target success probability throughout a D-step computation for every query in $\mathcal{U}$. If a query assigns at most $C/\sqrt{m}$ weight to any one of the m reached states, let $\mathcal{U}_C$ contain all queries that satisfy this bound. We prove

$$\mathfrak{D}(\boldsymbol{a}_{\mathrm{cum}};\mathcal{U}) \le C^2 \inf_{\boldsymbol{a}} \mathfrak{D}(\boldsymbol{a};\mathcal{U}), \qquad \boldsymbol{a}_{\mathrm{cum}} \in \arg\min_{\boldsymbol{a}} \mathfrak{D}(\boldsymbol{a};\mathcal{U}_C). \tag{1}$$

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

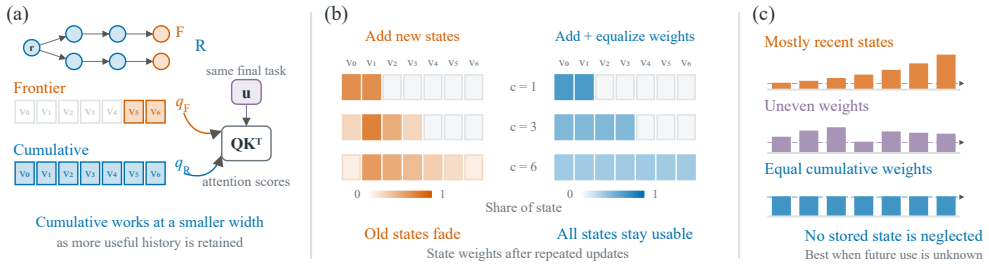


Figure 1: **The cumulative-state hypothesis from latent state to computation.** (a) The frontier state keeps only the states active now, whereas the cumulative state keeps all reached states for the same final task. (b) Each update adds new states; equalizing the weights keeps every reached state available. (c) Equal cumulative weights avoid a weakly represented state and are optimal when future use is unknown.

Thus, for a known query family, equal cumulative weights require at most C^2 times the hidden width of the best task-specific weights. When only the coefficient bound is known, equal cumulative weights are the best fixed choice in the worst case over the full class $\mathcal{U}_C$.

The query condition covers computations that combine information across a set, add costs along a route, execute an algorithm, or evaluate intermediate steps (Zaheer et al., 2017; Lee et al., 2019; Kool et al., 2019; Veličković et al., 2022; Lightman et al., 2024). We test the predictions with synthetic query families, paired width sweeps in a lightweight Transformer and a model built from GPT-2 blocks, and position-level interventions. Across these tests, the cumulative state reaches the target accuracy at smaller width, while unequal weights fail first at the states assigned the least weight.

Contributions. Our main contributions are:

1. **A counterintuitive result about latent-state capacity.** We show when preserving more reached states makes reasoning easier: their useful components reinforce later attention, allowing the cumulative state to work at a smaller hidden width than the frontier state.
2. **A Transformer account of the full reasoning process.** We track how attention retrieves states, values carry new information, the residual preserves earlier states, and the output projection sets their weights across all D steps. This analysis identifies the least represented state as the bottleneck for later computation.
3. **An optimal rule for weighting reached states.** Equal cumulative weights are optimal when future use is unknown and remain within a C^2 factor of the best task-specific weights when future queries are known. Two-layer and GPT-2 experiments support both the width advantage and the predicted failure of unequal weights.

2 Related Work

Explicit and continuous reasoning. Chain-of-thought writes intermediate computation as tokens (Wei et al., 2022; Merrill & Sabharwal, 2024), whereas COCONUT feeds latent states back into the model (Hao et al., 2024). Recent theory shows that a single continuous thought can preserve and expand several reasoning traces in parallel (Zhu et al., 2025b; Gozeten et al., 2025), and subsequent work studies how such superposition emerges during training (Zhu et al., 2025a). Zhu et al. (Zhu et al., 2025b) establish that frontier superposition can implement parallel search: the latent state keeps the alternatives active at the current step and expands them together. We build on this representation and study its capacity limit over a complete D-step Transformer computation. In particular, we compare keeping only the frontier with keeping all reached states, and we determine how the latent state should distribute weight when later queries are not known in advance.

Capacity of superposition. Work on feature superposition explains how non-orthogonal features share a low-dimensional space and how interference depends on sparsity, importance, and polysemanticity (Elhage et al., 2022; Foote, 2024; Gupta & Kumar, 2025; Bereska et al., 2025). These

analyses ask how many features can be represented and whether individual features remain recoverable. Our setting instead asks whether a normalized latent state can support a later attention computation. When that computation combines several reached states, their aligned components add coherently while unrelated candidates remain random interference. This produces the reversal studied in our theory: preserving more reached states can reduce the hidden width required for correct attention. We further compare all nonnegative state weightings and show when equal cumulative weights are optimal.

Recurrent computation and latent states. Universal and looped Transformers reuse computation depth while updating a latent state (Dehghani et al., 2019; Giannou et al., 2023; Xu & Sato, 2024; Fu et al., 2026). Recurrent and buffer-based models similarly carry information through dedicated latent tokens (Bulatov et al., 2022; Yang et al., 2024; Lin et al., 2026). Related work studies multi-hop reasoning in long contexts and without explicit intermediate steps (Li et al., 2024; Yao et al., 2025), as well as the limits and internal representations of hidden reasoning (Brown-Cohen et al., 2026; Hu et al., 2024). We connect these questions to the Transformer operations that update a latent state: attention retrieves earlier states, values carry new information, and the residual path determines what remains available. Our analysis identifies the least represented state as the bottleneck and yields a concrete principle for repeated computation: when future use is unknown, keep reached states equally weighted.

3 Problem Setting and Central Hypothesis

We study the continuous-thought token of a recurrent two-block Transformer. Following the copied-edge interface of Zhu et al. (2025b), object embeddings are available to attention as keys and values. A recurrent cell first uses $QK^\top$ attention to decide which objects are active, then writes their values into the next thought. We use this common cell to compare one proposed state construction against its natural baseline and against all nonnegative history weightings.

Let $G_{\text{graph}} = (V, E)$ be a directed graph with $n = |V|$ and root r. At a fixed reasoning depth, let

$$F := \{v : \operatorname{dist}(r, v) = c\}, \qquad R := \{v : \operatorname{dist}(r, v) \le c\}, \tag{2}$$

with $F \subseteq R$, $|F| = k$, and $|R| = m$, where $1 \le k \le m$. Vertices represent intermediate facts, entities, subgoals, partial plans, or candidate states. Across reasoning steps we write

$$R_c := \{v : \operatorname{dist}(r, v) \le c\}, \qquad F_c := \{v : \operatorname{dist}(r, v) = c\}, \qquad N_c := R_{c+1} \setminus R_c. \tag{3}$$

When k is the number of distinct states first reached at the current depth, whereas m is the number first reached up to that depth. Thus a length-c chain has $k = 1$ and $m = c + 1$.

Each vertex has an independent Gaussian embedding

$$g_v \sim N(0, I_d/d). \tag{4}$$

Equivalently, collect the embeddings as columns of $G \in \mathbb{R}^{d\times n}$ with independent $N(0, 1/d)$ entries. For a coefficient vector $a \in \mathbb{R}^n$, write

$$X_a := Ga = \sum_{v\in V} a_v g_v, \qquad q_a := \frac{X_a}{\|X_a\|}. \tag{5}$$

Multiplying a by a positive scalar does not change q_a, so we use $\|a\|_2 = 1$ throughout.

The proposed state, its baseline, and the comparison family are

$$a_F := \frac{\mathbf{1}_F}{\sqrt{k}}, \qquad a_R := \frac{\mathbf{1}_R}{\sqrt{m}}, \qquad a \in \mathcal{A}_R := \{a \in \mathbb{R}^n_+ : \operatorname{supp}(a) \subseteq R,\ \|a\|_2 = 1\}, \tag{6}$$

$$q_F := q_{a_F}, \qquad q_R := q_{a_R}, \qquad q_a := \frac{X_a}{\|X_a\|}. \tag{7}$$

Frontier is the baseline that retains only current alternatives. Uniform cumulative is the proposed state. The set $\mathcal{A}_R$ contains every nonnegative history weighting against which we compare that proposal.

Let $h(a)$ denote the thought token carrying q_a, and let h_j be a candidate token carrying coefficient direction u_j. The final Transformer head uses

$$W_Q h(a) = q_a, \qquad W_K h_j = X_{u_j} = G u_j, \qquad W_V h_j = v_j, \tag{8}$$

where, independently of G,

$$u_0, u_1, \ldots, u_M \in \mathbb{R}^n \tag{9}$$

are unit directions, u_0 is useful, $u_1, \ldots, u_M$ compete with it, and v_j is the value routed by candidate j. Its actual forward pass is

$$\alpha_j(q_a) := \frac{\exp(\langle q_a, X_{u_j}\rangle/\tau)}{\sum_{i=0}^{M} \exp(\langle q_a, X_{u_i}\rangle/\tau)}, \qquad \mathrm{Attn}(q_a) = \sum_{j=0}^{M} \alpha_j(q_a) v_j, \tag{10}$$

and the selected candidate is

$$\widehat{j}(q_a) := \arg\max_{0\le j\le M} \alpha_j(q_a) = \arg\max_{0\le j\le M} \langle q_a, X_{u_j}\rangle. \tag{11}$$

Success means that this head assigns its largest attention weight to the useful value, $\widehat{j}(q_a) = 0$. A node key is recovered by $u_0 = e_v$. The final attention head uses the same M competing keys for every state weighting. We assume $u_1, \ldots, u_M$ are orthonormal and orthogonal to $\mathrm{span}\{a, u_0\}$, so no competing key has a built-in alignment advantage. When comparing states supported on R, we choose the bank orthogonal to the R coordinates, making it common to all weightings and requiring $M \le n - m$. Appendix C treats a general fixed bank.

For the comparisons below, the useful direction u_0 is nonnegative and supported on R. The population quantity is

$$\mu(a, u_0) := \langle a, u_0\rangle \in [0, 1]. \tag{12}$$

Nonnegative a, u_0 express the positive state components represented by the superposition construction. Let $\mathcal{U}$ be the set of possible useful directions, all nonnegative, unit-normalized, and supported on R. We assume that no direction places more than $C/\sqrt{m}$ weight on one reached object:

$$\|u\|_\infty \le \frac{C}{\sqrt{m}}, \qquad u \in \mathcal{U}, \tag{13}$$

where $C \ge 1$. It measures how unevenly a future computation may use the reached history. A uniform history query has $C = 1$; bounded variation over a constant fraction of the history has $C = O(1)$.

Central hypothesis. Among all sequences of nonnegative history weights used by the recurrent Transformer, uniform cumulative's joint width measure should be within C^2 of the best sequence chosen with advance knowledge of $\mathcal{U}$. If only condition (13) is known and the realized future query is not, uniform cumulative should be the best single choice. Sections 4 and 5 prove these statements; Section 6 tests their predicted ordering and Transformer mechanism.

4 The Width of a D-Step Transformer Computation

We measure the width needed to keep a state usable throughout D attention calls and the final answer. Fix a coefficient sequence $\boldsymbol{a} = (a_c)_{c=0}^{D}$, independently of G, with $a_c \in \mathcal{A}_{R_c}$. Let S_c contain the sources that must remain usable at step c, and let $I_c \subseteq V \setminus \mathrm{supp}(a_c)$ contain the inactive sources. Define

$$a_{\min,c} := \min_{v\in S_c} a_{c,v}, \qquad \alpha_\star := \min_{c<D} a_{\min,c}, \qquad \tau_c := 3a_{\min,c}/8. \tag{14}$$

Frontier uses $S_c = F_c$; full reuse of cumulative states uses $S_c = R_c$. The intended states q_{a_c} give the joint routing event

$$\mathcal{E}_{\mathrm{route}}(d) := \bigcap_{c<D} \left\{ \min_{v\in S_c} \langle q_{a_c}, g_v\rangle > \tau_c > \max_{s\in I_c} \langle q_{a_c}, g_s\rangle \right\}. \tag{15}$$

For a nonempty compact final query class $\mathcal{U} \subseteq \mathcal{A}_{R_D}$, put

$$\Gamma(a_D; \mathcal{U}) := \inf_{u\in\mathcal{U}} \langle a_D, u\rangle. \tag{16}$$

Write d_D for the least width at which the intended states satisfy both (15) and the correct final-answer event with probability at least $1-\delta$ for every fixed $u \in \mathcal{U}$. Its width scale is

$$\mathfrak{D}(\boldsymbol{a};\mathcal{U}) := \max\left\{\max_{c<D} \frac{\log(Dn/\delta)}{a^2_{\min,c}}, \frac{2\log M}{\Gamma(a_D;\mathcal{U})^2}\right\}. \tag{17}$$

The actual Transformer state is denoted by $\widetilde{q}_c$. The construction in Section 5.3 receives copied edge buffers, fixed masks, deduplicated new destinations, and their counts. These inputs and all target coefficients are fixed before G is sampled. The following result connects the width comparison to repeated attention, residual writing, and RMSNorm.

Theorem 4.1 (Width and stability of D-step Transformer computation)**.** *Fix D and $\delta \in (0,1)$, assume $\alpha_\star > 0$ and $\Gamma := \Gamma(a_D;\mathcal{U}) > 0$, and use $\Theta(n)$ inactive node keys at each step and the common final bank. As $M \to \infty$, with n large enough for the bank, the intended-state joint requirement satisfies*

$$d_D(\boldsymbol{a};\mathcal{U},M,\delta) = \Theta(\mathfrak{D}(\boldsymbol{a};\mathcal{U}))\,. \tag{18}$$

In particular, the routing width and final-answer width obey

$$d_{\mathrm{route}} = \Theta\left(\max_{c<D} \frac{\log(Dn/\delta)}{a^2_{\min,c}}\right), \tag{19}$$

$$d_{\mathrm{ans}} = (2+o(1))\frac{\log M}{\Gamma^2}. \tag{20}$$

For the cumulative or frontier updates in Section 5.3, there are finite attention scales and a constant $K_{D,\delta}$ such that $d \geq K_{D,\delta}\mathfrak{D}(\boldsymbol{a};\mathcal{U})$ makes the actual D-step computation and final answer correct with probability at least $1-\delta$, for every $u \in \mathcal{U}$. On the same event,

$$\max_{c\leq D} \|\widetilde{q}_c - q_{a_c}\| \leq A_D\left(\sqrt{\frac{\log(16(D+1)/\delta)}{d}} + \zeta\right), \tag{21}$$

where ζ bounds each destination-write error and can be made arbitrarily small by finite attention scales. Constants may depend on the fixed number D of recurrent steps.

Remark 4.2. *A state must support both the next attention call and the eventual answer. The smallest retained coefficient controls the first requirement, while alignment with the final query controls the second. The stability bound shows how the same width scale also supports repeated Transformer updates: writing and normalization errors stay below the attention gaps across the whole trace, rather than being reset after each step.*

5 Theoretical Validation of Cumulative State

We now compare complete D-step computations through the width scale in Theorem 4.1. The comparison keeps the graph, Transformer cell, and final candidate bank fixed; only the coefficients carried by the thought token change.

5.1 The C^2 guarantee for a known query class

The query condition in Section 3 is equivalently $\mathcal{U} \subseteq \mathcal{U}_C(R)$, where

$$\mathcal{U}_C(R) := \left\{u \in \mathcal{A}_R : \|u\|_\infty \leq C/\sqrt{|R|}\right\}; \tag{22}$$

the class may be finite or continuous and need not treat history positions symmetrically.

Theorem 5.1 (Cumulative guarantee and best fixed weighting)**.** *Suppose every reached object remains eligible for later source-membership attention. Let $\mathcal{U} \subseteq \mathcal{U}_C(R_D)$ be any nonempty compact query class, and let $\boldsymbol{a}_R = (a_{R_c})_{c=0}^D$ be the uniform cumulative sequence. Then*

$$\mathfrak{D}(\boldsymbol{a}_R;\mathcal{U}) \leq C^2 \inf_{\substack{a_c\in\mathcal{A}_{R_c}\\ 0\leq c\leq D}} \mathfrak{D}(\boldsymbol{a};\mathcal{U}). \tag{23}$$

The factor C^2 for the width scale $\mathfrak{D}$ cannot in general be reduced. If the realized query is not known beyond the condition defining $\mathcal{U}_C(R_D)$, cumulative is the best single weighting for the full allowed family:

$$\boldsymbol{a}_R \in \arg \min_{\substack{a_c \in \mathcal{A}_{R_c} \\ 0 \le c \le D}} \mathfrak{D}(\boldsymbol{a}; \mathcal{U}_C(R_D)). \tag{24}$$

Remark 5.2. *The comparison allows every nonnegative coefficient sequence, including sequences that a scalar residual update cannot generate. Equal cumulative weights still lose at most C^2 against the best such choice for a known query class, and minimize the full-family objective when the query is unknown. The same inequalities hold over any smaller comparison family containing the cumulative sequence. Together with Theorem 4.1, this gives an actual cumulative implementation at width at most $K_{D,\delta} C^2 \inf_{\boldsymbol{a}} \mathfrak{D}(\boldsymbol{a}; \mathcal{U})$: the optimal weight comparison and the stability of its Transformer implementation are both controlled.*

5.2 Why preserving history can reduce width

The query condition in this subsection reads

$$\|u\|_\infty \le \frac{C}{\sqrt{m}}, \qquad u \in \mathcal{U}. \tag{25}$$

Tasks covered by the condition. Since $\|u\|_2 = 1$, a fixed C means that every query draws on at least m/C^2 reached objects. Uniform history readouts have $C = 1$, while bounded-imbalance aggregation over a constant fraction has $C = O(1)$. This covers whole-set aggregation, additive objectives over visited objects, and supervision over algorithmic traces (Zaheer et al., 2017; Lee et al., 2019; Kool et al., 2019; Veličković et al., 2022; Lightman et al., 2024). The regime $k = o(m)$ includes chains, fixed-width traces, and grid-like breadth-first search: the current frontier grows more slowly than the latent states back into computation.

Theorem 5.3 (When cumulative reduces the width of the complete trace)**.** *Fix $F \subsetneq R$, with $|F| = k < m = |R|$, and use the same candidate bank for both states. Let $\mathcal{U} \subseteq \mathcal{A}_R$ be any nonempty compact query class satisfying* (25)*. Then*

$$\Gamma(a_R; \mathcal{U}) \ge \frac{1}{C}, \qquad \Gamma(a_F; \mathcal{U}) \le C\sqrt{\frac{k}{m}}, \tag{26}$$

Let

$$T_R := \frac{2 \log M}{\Gamma(a_R; \mathcal{U})^2}, \qquad T_F := \frac{2 \log M}{\Gamma(a_F; \mathcal{U})^2} \tag{27}$$

be the final-answer terms of the two D-step computations. Whenever $\Gamma(a_F; \mathcal{U}) > 0$,

$$\frac{T_R}{T_F} \le C^4 \frac{k}{m}. \tag{28}$$

Writing R_R, R_F for their routing terms in (17)*, their complete width scales are $\max\{R_R, T_R\}$ and $\max\{R_F, T_F\}$. Hence the same ratio describes the end-to-end dimensions whenever final answer selection is the active bottleneck for both policies. For fixed C it then vanishes in the growing-history regime $k = o(m)$. The rate $C^4 k/m$ is sharp: it is attained when $k \le m/C^2$ and m/C^2 is an integer, and asymptotically attained along fixed-C growing-history sequences with $m/C^2 \to \infty$. If $\Gamma(a_F; \mathcal{U}) = 0$, frontier's final head cannot exceed chance at any finite dimension while cumulative still can.*

Remark 5.4. *This theorem identifies when the end-to-end gain comes from the final use of state. As the reached history grows, cumulative strengthens the correct answer logit even though it stores more objects. Equation* (17) *also shows when intermediate routing, rather than final readout, sets the required width; the C^2 result above covers both bottlenecks simultaneously.*

For the representative history-wide direction $u = a_R$, the final terms are exactly

$$T_F = \Big(2\frac{m}{k} + o\Big(\frac{m}{k}\Big)\Big) \log M, \qquad T_R = (2 + o(1)) \log M. \tag{29}$$

Thus the cumulative final state uses a factor m/k less width for this history-wide computation; Theorem 4.1 combines this gain with the cost of producing all preceding states.

The reason is coherent addition. Readout succeeds when

$$\underbrace{\langle a, u\rangle\sqrt{d}}_{\text{useful signal}} \gtrsim \underbrace{\sqrt{2\log M}}_{\text{unrelated competition}} . \tag{30}$$

For cumulative, $\langle a_R, u\rangle = \|u\|_1/\sqrt{m}$: useful historical components add before normalization, while unrelated candidates remain centered fluctuations. The extra stored objects therefore strengthen a history-wide computation instead of acting as independent noise. This is the mathematical source of the apparent reversal.

5.3 How the same Transformer writes frontier or cumulative

Each recurrent call receives the thought, a threshold token, and copied source–destination buffers. The fixed interface supplies one active edge per newly reached destination in N_c, removes already reached destinations, and supplies $m_c = |R_c|$ and $r_c = |N_c|$. Thus destinations are counted once; every active source lies in S_c and every inactive source lies in I_c. These masks, copies, and counts are inputs to the construction. We consider $r_c \geq 1$; a trace stops when no new destination remains.

For an edge $e = (s, t)$, the first block uses

$$Q_1 h_e = [g_s; 1], \qquad K_1 h_c = [\widetilde{q}_c; 0], \qquad K_1 h_{\tau_c} = [0; \tau_c]. \tag{31}$$

The edge attends only to the thought and threshold tokens. Their values 1 and 0 give the gate

$$z_e = \sigma(\beta(\langle \widetilde{q}_c, g_s\rangle - \tau_c)) . \tag{32}$$

The second block reads the gate and copies destination values:

$$Q_2 h_{c+1} = 1, \qquad K_2 h_e = T(2z_e - 1), \qquad V_2 h_e = g_t. \tag{33}$$

After multiplying its attention average by $\sqrt{r_c}$, its output satisfies

$$\widetilde{o}_c = \frac{1}{\sqrt{r_c}} \sum_{t\in N_c} g_t + \xi_c, \qquad \|\xi_c\| \leq \zeta. \tag{34}$$

A masked copy head returns $\widetilde{q}_c$. Prescribed scalar gates λ_c, η_c scale the copy and destination branches; a fixed output projection combines them with the residual. RMSNorm, with gain chosen to return unit state norm, gives

$$\widetilde{q}_{c+1} = \mathrm{RMSNorm}(\lambda_c \widetilde{q}_c + \eta_c \widetilde{o}_c). \tag{35}$$

Here d is the node/key channel width; a constant number of buffer channels uses $O(d)$ hidden width. The scalar controls include the supplied counts. The two choices are

$$(\lambda_c, \eta_c) = (0, 1) \quad \text{(frontier)}, \qquad (\lambda_c, \eta_c) = (\sqrt{m_c}, \sqrt{r_c}) \quad \text{(cumulative)}. \tag{36}$$

Starting from the normalized root state, the latter approximates uniform cumulative states with the error in (21). Other scalar gates produce weights by arrival batch: they rescale all retained coefficients together and assign a common coefficient to the new batch. The comparison in Theorem 5.1 allows the larger family of all nonnegative weight sequences. Cumulative's guarantee therefore holds even against choices beyond those generated by these scalar gates.

6 Experimental Evidence

We test the cumulative-superposition hypothesis at the same three levels as the theory. Within every comparison, the task, candidates, embedding codebook, and trained attention mechanism are shared; only the imposed state coefficients change.

6.1 Test I: Does Cumulative Need Less Width?

We test Theorem 5.3 on a grid over $m \in \{64, 128, 256, 512\}$, $k/m \in \{1/32, \ldots, 1/2\}$, $C \in \{1, 1.25, 1.5, 2\}$, and $M \in \{2^{10}, 2^{14}\}$, estimating worst-case $D_\star$ by binary search on d with 2000 Monte Carlo trials ($\delta = 0.1$). Each query is drawn independently of G and projected onto the cap $\|u\|_\infty \leq C/\sqrt{m}$. *Dense* uses random positive weights on all m coordinates, modeling uneven aggregation over a visited set (Zaheer et al., 2017). *Temporal* centers an exponential bump along the trace, modeling localized process supervision over a CoT (Lightman et al., 2024). *Block* partitions the trace into segments with different scales, modeling staged algorithmic readouts (Veličković et al., 2022). *Union* mixes the three generators. The *sharp witness* is the theorem's example, uniform on $\lceil m/C^2 \rceil$ coordinates that contain F. *Zero-overlap* is uniform on $R \setminus F$, so $\Gamma(a_F; \mathcal{U}) = 0$.

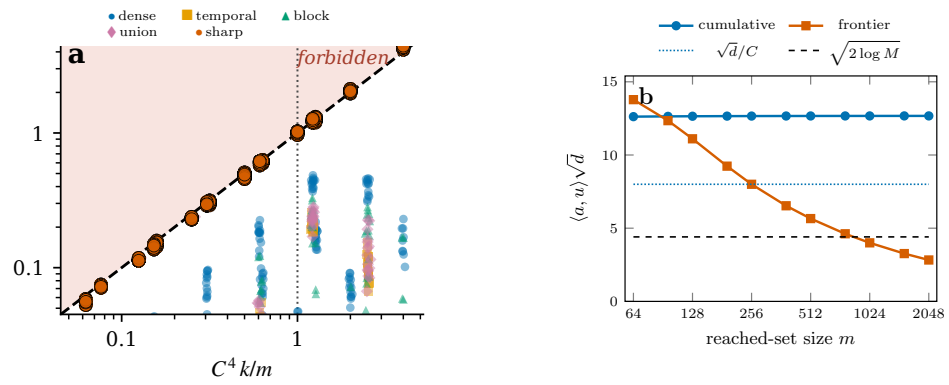
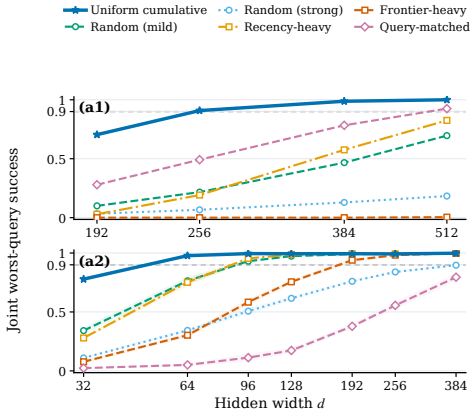


Figure 2: **Why remembering more can require less width.** (a) Across many ways of using the reached history state, cumulative state stays below the predicted width bound, while the deliberately hardest examples approach it. (b) As the reached history grows, cumulative keeps useful signal for the final task, whereas frontier discards earlier states and becomes harder to use.



(b) Paired five-seed results

Lightweight two-block model			
State	Acc. @ 32	Acc. @ 40	Empirical $d_{0.9}$
Frontier	0.210 ± 0.007	0.322 ± 0.005	110.6 ± 1.6
Cumulative	$\mathbf{0.793 \pm 0.015}$	$\mathbf{0.899 \pm 0.010}$	$\mathbf{40.3 \pm 1.1}$
Ideal frontier	0.216 ± 0.007	0.328 ± 0.006	112
Ideal cumulative	0.947 ± 0.005	0.990 ± 0.002	32
GPT-2 backbone			
State	Acc. @ 32	Acc. @ 64	Empirical $d_{0.9}$
Frontier	0.468 ± 0.038	0.900 ± 0.026	66.2 ± 5.9
Cumulative	$\mathbf{0.898 \pm 0.024}$	$\mathbf{0.998 \pm 0.003}$	$\mathbf{34.4 \pm 3.8}$

Width ratio $d_{0.9}^{\mathrm{C}}/d_{0.9}^{\mathrm{F}} \downarrow$	Two-layer	GPT-2
Paired estimate	0.365 ± 0.008	0.521 ± 0.067

Figure 3: **Trained Transformers reproduce the predicted width advantage.** The shared legend shows six ways to weight the same reached history. (a1) Results for the two-layer block model; (a2) results for the GPT-2-block model, with bands showing variation across runs. Every curve in a panel uses the same trained attention module, and equal cumulative weighting reaches 90% accuracy on the hardest tested task first. (b) Five paired runs confirm that cumulative reaches the same held-out accuracy at a smaller width than frontier in both architectures. Here $d_{0.9}$ means the smallest interpolated width reaching 90% accuracy; Ideal supplies the intended state directly.

Protocol. We use the ProsQA(Zhong et al., 2022) graph-state task and hold the graph trace, candidate bank, and Gaussian codebook fixed when comparing state constructions. Queries are drawn before the Gaussian map. The controlled attention experiment uses an identity-regularized source-membership router at $d = 192$ and evaluates five independent Gaussian codebooks.

6.2 Test II: Do Trained Transformers Show the Same Advantage?

Across the paired five-seed sweeps, cumulative crosses the same 90% held-out success threshold at substantially lower width on both architectures. Its interpolated minimum reliable width is 40.3 ± 1.1, compared with 110.6 ± 1.6 for frontier, in the lightweight two-block model; with the two-layer GPT-2 backbone it is 34.4 ± 3.8, compared with 66.2 ± 5.9. The corresponding cumulative/frontier ratios are 0.365 ± 0.008 and 0.521 ± 0.067. Thus the separation predicted by Theorem 5.3 persists when the analytic state construction is replaced by a trained causal Transformer. Figure 3 shows both width sweeps and their paired minimum-width estimates under matched evaluation protocols.

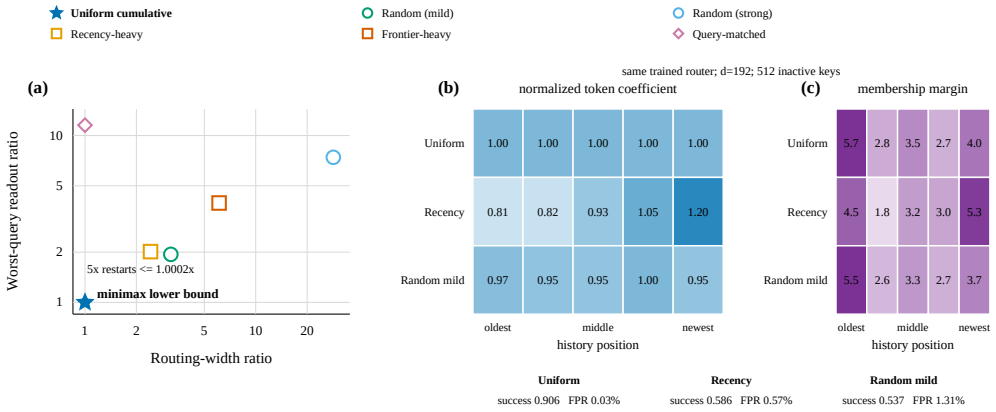


Figure 4: **Why equal cumulative weights are safest.** (a) Each point compares the width needed to keep states usable during reasoning and the width needed to use them for the final answer; lower is better. (b) The heatmap shows how much state is assigned to each position in the reached history. (c) Using the same trained attention module, states with small weights are the first to be missed. The summary reports success on the whole routing step and the rate at which inactive states are selected by mistake.

6.3 Test III: What Fails Under Unequal Weights?

We next test the mechanism behind Theorem 5.1. Figure 4(a) shows the predicted two-term geometry: uniform cumulative is the joint lower-left point after normalizing routing and worst-query readout width by the uniform construction. Panels (b,c) expose what this optimum means inside the router. Uniform weights equalize the coefficient of every reusable state token; recency and random weights produce low-coefficient tokens, and those tokens receive lower membership margins. Consequently, joint route success falls from 0.906 to 0.586 and 0.537, while the per-inactive-key false-positive rate rises from 0.03% to 0.57% and 1.31%, respectively.

7 Conclusion

We asked how reasoning should be organized inside a Transformer's latent space. Our answer is the cumulative-superposition hypothesis: when past reasoning may be useful again, the latent state should preserve it with equal weights instead of keeping only the latest frontier.

The theory explains why this can help. Useful parts of the history reinforce one another in later attention, so cumulative state can reach reliable choices at a smaller hidden width than frontier state. This advantage can survive multi-step Transformer computation: attention finds relevant states, values carry new information, the residual keeps earlier information, and the feed-forward layer restores equal weights after each write. Finally, when the model does not know which part of state a future task will use, equal weights give the strongest guarantee; when the task family is known, they remain within the explicit C^2 factor of the best specialized weighting.

The experiments match these predictions. Both the lightweight and GPT-2-block Transformers reach the same accuracy at smaller width with cumulative state. The position-level tests also show the predicted mechanism: unequally weighted states fail first where they assign the least weight.

Together, the results turn cumulative superposition into a concrete design principle for continuous reasoning: preserve the reached states evenly and restore that balance as computation proceeds. This allows a fixed hidden width to support longer reusable histories and more reliable future computation.

AI USE STATEMENT

Generative AI tools were only used for grammar correction and writing optimization, as well as generating some chart-drawing code, and all AI-assisted content was carefully reviewed by the author. The author takes full responsibility for the final content of this work, including its statements, proofs, experiments, citations, charts, and text.

REFERENCES

Leonard Bereska, Zoe Tzifa-Kratira, Reza Samavi, and Efstratios Gavves. Superposition as lossy compression: Measure with sparse autoencoders and connect to adversarial vulnerability. *arXiv preprint arXiv:2512.13568*, 2025.

Jonah Brown-Cohen, David Lindner, and Rohin Shah. Quantifying the necessity of chain of thought through opaque serial depth. *arXiv preprint arXiv:2603.09786*, 2026.

Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

Alex Foote. Tackling polysemanticity with neuron embeddings. *arXiv preprint arXiv:2411.08166*, 2024.

Rao Fu, Zixuan Yang, Jiankun Zhang, Jing Ma, Hechang Chen, Yu Li, and Yi Chang. Simply stabilizing the loop via fully looped transformer. *arXiv preprint arXiv:2605.18797*, 2026.

Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11398–11442, 2023.

Halil Alperen Gozeten, M. Emrullah Ildiz, Xuechen Zhang, Hrayr Harutyunyan, Ankit Singh Rawat, and Samet Oymak. Continuous chain of thought enables parallel exploration and reasoning. *arXiv preprint arXiv:2505.23648*, 2025.

Manan Gupta and Dhruv Kumar. Disentangling polysemantic neurons with a null-calibrated polysemanticity index and causal patch interventions. *arXiv preprint arXiv:2508.16950*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Lijie Hu, Liang Liu, Shu Yang, Xin Chen, Zhen Tan, Muhammad Asif Ali, Mengdi Li, and Di Wang. Understanding reasoning in chain-of-thought from the hopfieldian view. *arXiv preprint arXiv:2410.03595*, 2024.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2019.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753, 2019.

Yanyang Li, Shuo Liang, Michael R. Lyu, and Liwei Wang. Making long-context language models better multi-hop reasoners. *arXiv preprint arXiv:2408.03246*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Hanyu Lin, Min Cai, Jiawei Wen, and Haodi Zhang. Tyler: Typed latent reasoning for language models – when to think, what to compute, and how much to allocate. *arXiv preprint arXiv:2606.16360*, 2026.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, 2024.

Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, and Charles Blundell. The CLRS algorithmic reasoning benchmark. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22084–22102, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kevin Xu and Issei Sato. On expressive power of looped transformers: Theoretical analysis and enhancement via timestep encoding. *arXiv preprint arXiv:2410.01405*, 2024.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *arXiv preprint arXiv:2406.04271*, 2024.

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. *arXiv preprint arXiv:2505.17923*, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. ProQA: Structural prompt-based pre-training for unified question answering. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4230–4243, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.313. URL https://aclanthology.org/2022.naacl-main.313/.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Emergence of superposition: Unveiling the training dynamics of chain of continuous thought. *arXiv preprint arXiv:2509.23365*, 2025a.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. In *Advances in Neural Information Processing Systems*, 2025b. arXiv:2505.12514.

A Proofs for the Dimension Limit

The complete-trace theorem uses the following local attention facts. They are stated here because an individual call is a proof component rather than the object compared in the main text.

For this appendix only, let

$$d_\star(a, u_0; M, \delta) := \min\{d : \mathbb{P}(\widehat{j}(q_a) = 0) \geq 1 - \delta\}, \tag{37}$$

with $d_\star = \infty$ when $\mu(a, u_0) = 0$ and $1 - \delta > 1/(M + 1)$.

Lemma A.1 (Exact distribution of one attention comparison). *Fix unit a, u_0 with $\mu := \langle a, u_0 \rangle \in [0,1]$, and assume* (3)*. There exist mutually independent $Z_0, Z_1, \ldots, Z_M \sim N(0,1)$ and $C_d \sim \chi_d$ such that*

$$\sqrt{d}\,\langle q_a, X_{u_0} \rangle \overset{d}{=} \mu C_d + \sqrt{1-\mu^2}\, Z_0, \tag{38}$$

$$\sqrt{d}\,\langle q_a, X_{u_j} \rangle \overset{d}{=} Z_j, \qquad 1 \le j \le M. \tag{39}$$

Consequently,

$$\mathbb{P}\left(\alpha_0(q_a) = \max_{0 \le j \le M} \alpha_j(q_a)\right) = \mathbb{P}\left(\mu C_d + \sqrt{1-\mu^2}\, Z_0 > \max_{1 \le j \le M} Z_j\right). \tag{40}$$

Lemma A.2 (Monotonicity in alignment). *For fixed $d \ge 1$ and $M \ge 1$, the probability in* (40) *is strictly increasing in $\mu \in [0,1]$.*

Theorem A.3 (Local attention transition). *If $M \to \infty$ and $d_M \mu_M^2 / (2 \log M) \to \lambda$, then*

$$\mathbb{P}\left(\alpha_0(q_a) = \max_{0 \le j \le M} \alpha_j(q_a)\right) \longrightarrow \begin{cases} 0, & \lambda < 1, \\ 1, & \lambda > 1. \end{cases} \tag{41}$$

Corollary A.4 (Dimension of one attention comparison). *For fixed $\delta \in (0,1)$,*

$$d_\star(a, u_0; M, \delta) = (2 + o(1)) \frac{\log M}{\mu_M^2}. \tag{42}$$

Proof of Lemma A.1. Let $Y := Ga$. Since $\|a\|_2 = 1$, $Y \sim N(0, I_d/d)$ and $\sqrt{d}\,\|Y\| \sim \chi_d$.

If $\mu < 1$, define

$$c := \frac{u_0 - \mu a}{\sqrt{1-\mu^2}}. \tag{43}$$

Then c is unit norm, $c \perp a$, and $u_0 = \mu a + \sqrt{1-\mu^2} c$. If $\mu = 1$, the term involving c below is absent. Gaussianity and $c \perp a$ imply that Gc is independent of Y. Conditioned on Y, rotational invariance gives

$$\sqrt{d}\langle Y/\|Y\|, Gc \rangle \sim N(0,1), \tag{44}$$

and this random variable is independent of $\|Y\|$. Therefore

$$\begin{aligned} \sqrt{d}\langle q_a, G u_0 \rangle &= \mu \sqrt{d}\,\|Y\| + \sqrt{1-\mu^2}\sqrt{d}\langle Y/\|Y\|, Gc \rangle \\ &\overset{d}{=} \mu C_d + \sqrt{1-\mu^2} Z_0. \end{aligned} \tag{45}$$

For each $j \ge 1$, $u_j \perp a$, so Gu_j is independent of Y and

$$\sqrt{d}\langle Y/\|Y\|, Gu_j \rangle \sim N(0,1). \tag{46}$$

The orthogonality of $u_1, \ldots, u_M$ and their orthogonality to u_0 imply joint independence of all the Gaussian terms. Comparing the scores proves (40). □

Proof of Lemma A.2. Let $S_\mu := \mu C_d + \sqrt{1-\mu^2} Z$ and $F_\mu(t) := \mathbb{P}(S_\mu \le t)$. For $0 < \mu < 1$, put $\sigma^2 := 1 - \mu^2$. Differentiating under the integral against the χ_d density gives

$$\frac{\partial F_\mu(t)}{\partial \mu} = -K_{d,\mu,t} \int_0^\infty (c - \mu t) c^{d-1} \exp\left(-\frac{(c-\mu t)^2}{2\sigma^2}\right) dc, \tag{47}$$

where $K_{d,\mu,t} > 0$. Denote the integral by I. Since $(c-\mu t) e^{-(c-\mu t)^2/(2\sigma^2)}$ equals $-\sigma^2$ times the derivative of the exponential, integration by parts yields

$$I = \begin{cases} \sigma^2 \exp\left(-\dfrac{\mu^2 t^2}{2\sigma^2}\right), & d = 1, \\ \sigma^2 (d-1) \displaystyle\int_0^\infty c^{d-2} \exp\left(-\frac{(c-\mu t)^2}{2\sigma^2}\right) dc, & d > 1. \end{cases} \tag{48}$$

Thus $I > 0$ and $\partial_\mu F_\mu(t) < 0$ for every t. Continuity handles $\mu = 0, 1$, so S_μ is strictly increasing in first-order stochastic order. Conditional on $S_\mu = s$, success against M independent standard-normal decoys has probability $\Phi(s)^M$, a strictly increasing function of s. Therefore the unconditional success probability is strictly increasing in μ. The critical-dimension conclusion follows immediately. □

Lemma A.5 (Gaussian maximum and chi concentration)**.** *Let $Z_1, \dots, Z_M$ be independent $N(0,1)$ variables and let $C_d \sim \chi_d$. As $M, d \to \infty$,*

$$\frac{\max_{j \le M} Z_j}{\sqrt{2 \log M}} \to 1, \qquad \frac{C_d}{\sqrt{d}} \to 1 \tag{49}$$

in probability.

Proof. For the upper Gaussian bound, for every $\varepsilon > 0$, the union bound and $\mathbb{P}(Z \ge t) \le e^{-t^2/2}$ give

$$\begin{aligned} \mathbb{P}\left(\max_{j \le M} Z_j > (1+\varepsilon)\sqrt{2 \log M} \right) &\le M \exp(-(1+\varepsilon)^2 \log M) \\ &= M^{-2\varepsilon - \varepsilon^2} \to 0. \end{aligned} \tag{50}$$

For the lower bound, let $t = (1-\varepsilon)\sqrt{2 \log M}$. The Gaussian Mills lower bound gives, for a universal $c > 0$ and all sufficiently large t,

$$\mathbb{P}(Z > t) \ge c \frac{e^{-t^2/2}}{t}. \tag{51}$$

Therefore $M\mathbb{P}(Z > t) \to \infty$, and independence yields

$$\mathbb{P}\left(\max_{j \le M} Z_j \le t \right) = (1 - \mathbb{P}(Z > t))^M \le e^{-M\mathbb{P}(Z>t)} \to 0. \tag{52}$$

This proves the first convergence. For the second, C_d^2/d is the average of d independent squared standard Gaussians, so it converges to one in probability by the weak law of large numbers. The square root is continuous on $(0, \infty)$. □

Proof of Theorem A.3. Divide the two sides of (40) by $\sqrt{2 \log M}$. Lemma A.5 gives

$$\frac{\max_{j \le M} Z_j}{\sqrt{2 \log M}} \to 1. \tag{53}$$

If $\lambda = 0$, then

$$\mathbb{E}\left[\left(\frac{\mu_M C_{d_M}}{\sqrt{2 \log M}} \right)^2 \right] = \frac{d_M \mu_M^2}{2 \log M} \to 0, \tag{54}$$

so the aligned term converges to zero in probability without requiring $d_M \to \infty$.

Now suppose $\lambda \in (0, \infty]$. Since $\mu_M \le 1$, the assumed ratio implies $d_M \to \infty$, and The aligned term satisfies

$$\frac{\mu_M C_{d_M}}{\sqrt{2 \log M}} = \sqrt{\frac{d_M \mu_M^2}{2 \log M}} \frac{C_{d_M}}{\sqrt{d_M}} \to \sqrt{\lambda}, \tag{55}$$

where convergence to $+\infty$ is understood when $\lambda = \infty$. Because $|\sqrt{1 - \mu_M^2}| \le 1$ and a standard Gaussian is tight,

$$\frac{\sqrt{1 - \mu_M^2} Z_0}{\sqrt{2 \log M}} \to 0 \tag{56}$$

in probability. If $1 < \lambda < \infty$, choose a constant strictly between 1 and $\sqrt{\lambda}$; the true score lies above that constant and the maximum decoy below it with probability tending to one. If $0 < \lambda < 1$, choose a constant strictly between $\sqrt{\lambda}$ and 1 and reverse the argument. The cases $\lambda = 0$ and $\lambda = \infty$ follow from the two endpoint limits above. This proves (41). □

Additional consequence for the number of candidates. Let $N := n - m$ and define

$$M_{\max}(d, \mu, \delta, N) := \max\{1 \le M \le N : p_{d,M}(\mu) \ge 1 - \delta\}, \tag{57}$$

where $p_{d,M}(\mu)$ is the probability in Equation (40). If $\min\{d\mu^2, \log N\} \to \infty$, then

$$\log M_{\max} = (1 + o(1)) \min\left\{\frac{d\mu^2}{2}, \log N\right\}. \tag{58}$$

Proof of Corollary A.4. For fixed M and $\mu > 0$, the success probability in (40) is nondecreasing in d. To see this, couple $C_{d+1} = \sqrt{C_d^2 + Z^2}$ with an additional independent standard Gaussian Z; then $C_{d+1} \ge C_d$ almost surely, while all other variables in (40) are unchanged.

Fix $\varepsilon > 0$. Theorem A.3 implies that at dimensions

$$d_- = (1 - \varepsilon)\frac{2 \log M}{\mu_M^2}, \qquad d_+ = (1 + \varepsilon)\frac{2 \log M}{\mu_M^2}, \tag{59}$$

the success probabilities tend to zero and one, respectively. For all sufficiently large M, the critical dimension lies between d_- and d_+. Letting $\varepsilon \downarrow 0$ proves (42). Solving the same transition relation for $\log M$ gives (58). □

B Proofs for Frontier, Cumulative, and Weighted States

Proof of Theorem 5.3. Fix $u \in \mathcal{U}$. Nonnegativity, unit norm, and (25) give

$$1 = \|u\|_2^2 \le \|u\|_\infty \|u\|_1 \le \frac{C}{\sqrt{m}} \|u\|_1 . \tag{60}$$

Therefore

$$\langle a_R, u\rangle = \frac{\|u\|_1}{\sqrt{m}} \ge \frac{1}{C}. \tag{61}$$

For frontier,

$$\langle a_F, u\rangle = \frac{1}{\sqrt{k}} \sum_{v \in F} u_v \le \sqrt{k}\, \|u\|_\infty \le C\sqrt{\frac{k}{m}}. \tag{62}$$

Taking infima proves (26).

Compactness and Lemma A.2 imply that the worst-query success probability under state a is exactly $p_{d,M}(\Gamma(a;\mathcal{U}))$. If $\Gamma(a_F;\mathcal{U}) > 0$, the uniform critical-dimension law gives

$$\frac{D_\star(a_R;\mathcal{U}, M, \delta)}{D_\star(a_F;\mathcal{U}, M, \delta)} = \frac{\Gamma(a_F;\mathcal{U})^2}{\Gamma(a_R;\mathcal{U})^2}(1 + o(1)) \le C^4 \frac{k}{m}(1 + o(1)). \tag{63}$$

The remainder is uniform over positive alignments by the asymptotic convention. If $\Gamma(a_F;\mathcal{U}) = 0$, compactness supplies a query with zero frontier alignment; its correct answer is exchangeable with the M decoys, giving the stated infinite critical dimension.

It remains to prove sharpness. Suppose first that $s = m/C^2$ is an integer and $k \le s$. Choose $S \subseteq R$ with $|S| = s$ and $F \subseteq S$, and let $u_v = C/\sqrt{m}$ on S and zero elsewhere. Then u is a nonnegative unit vector satisfying (25). For the compact singleton class $\mathcal{U} = \{u\}$,

$$\Gamma(a_R;\mathcal{U}) = \frac{1}{C}, \qquad \Gamma(a_F;\mathcal{U}) = C\sqrt{\frac{k}{m}}, \tag{64}$$

so equality holds in (28). If m/C^2 is not an integer, use $\lfloor m/C^2 \rfloor$ coordinates at the cap and one residual coordinate. The rounding changes the ratio by $1 + o(1)$ whenever $m/C^2 \to \infty$. The condition $k \le m/C^2$ holds eventually in every fixed-C growing-history regime with $k = o(m)$. □

Equation (29) follows by taking $u = a_R$: then $\langle a_R, u\rangle = 1$ and $\langle a_F, u\rangle = \sqrt{k/m}$, and substitution into Corollary A.4 gives the two dimensions.

Proof of Theorem 4.1. All target coefficients, graph sets, masks, and candidate directions below are fixed independently of G. We first bound attention gaps for these intended states, then transfer the gaps to the actual recurrent computation. The probability statement is uniform in the choice of a fixed $u \in \mathcal{U}$, not a simultaneous event over all queries.

Intended-state attention gaps and width. At step c, set $a = a_c$ and $\alpha = a_{\min,c}$. Gaussian regression gives, for $v \in S_c$,

$$\sqrt{d}\,\langle q_a, g_v\rangle \overset{d}{=} a_v C_d + \sqrt{1-a_v^2}\,Z_v, \qquad C_d = \sqrt{d}\,\|Ga\| \sim \chi_d, \tag{65}$$

where each Z_v is marginally standard normal. For $s \in I_c$,

$$\sqrt{d}\,\langle q_a, g_s\rangle \sim N(0,1). \tag{66}$$

These inactive scores are independent within the step and independent of the intended state and required-source scores. The required-source scores need not be mutually independent. On the event

$$C_d/\sqrt{d} \ge 3/4, \qquad \min_{v\in S_c} Z_v \ge -\alpha\sqrt{d}/4, \qquad \max_{s\in I_c} \sqrt{d}\langle q_a, g_s\rangle \le \alpha\sqrt{d}/4, \tag{67}$$

required scores are at least $\alpha/2$, inactive scores are at most $\alpha/4$, and the threshold $3\alpha/8$ has margin $\alpha/8$. Gaussian and chi-square tails imply, for absolute constants $A, b > 0$,

$$\mathbb{P}(\mathcal{E}_{\mathrm{route}}(d)^c) \le A \sum_{c<D} n \exp(-bd\,a_{\min,c}^2). \tag{68}$$

The same bound applies to failure of the stronger margin event just used. For a lower bound, successful routing requires the weakest source at each step to beat $\Theta(n)$ inactive keys. Its score law is Lemma A.1 with alignment $a_{\min,c}$, so Corollary A.4 gives $\Omega(\log n/a_{\min,c}^2)$. Since D, δ are fixed, these lower bounds and (68) prove (19).

For any fixed final query u, Lemma A.1 applies unconditionally to the deterministic coefficients a_D, u. Monotonicity in alignment and compactness give the worst-query alignment $\Gamma = \Gamma(a_D; \mathcal{U})$. Consequently

$$d_{\mathrm{ans}} = (2 + o(1))\frac{\log M}{\Gamma^2}. \tag{69}$$

Here and in (20), $M \to \infty$ at fixed failure probability, with the remainder uniform over positive alignments as in Corollary A.4. Allocating half the error to routing and half to the final answer proves the upper bound for d_D. Joint success requires each event separately, proving its lower bound and hence (18). No independence between steps is used.

A common event for the actual computation. Fix $u \in \mathcal{U}$ and consider either the cumulative or frontier target sequence. Set $\rho = \min(\alpha_\star, \Gamma)$. For all $c \le D$ require

$$\big|\,\|Ga_c\| - 1\big| \le \epsilon, \qquad \epsilon = K\sqrt{\log(16(D+1)/\delta)/d} \le 1/4. \tag{70}$$

Also require every node key, Gu, and every competing final key to have norm at most 2. Gaussian norm concentration and a union bound ensure these events with total failure probability at most $\delta/2$, provided $d \ge K\log((n+M+D+1)/\delta)$ and K is sufficiently large. Assign probability $\delta/4$ to failure of the route-margin events above. For the final head, with $\mu = \langle a_D, u\rangle \ge \Gamma$, require

$$\langle q_{a_D}, Gu\rangle \ge \Gamma/2, \qquad \max_{j\ge 1}\langle q_{a_D}, Gu_j\rangle \le \Gamma/4. \tag{71}$$

Indeed, in the exact score law, $C_d/\sqrt{d} \ge 3/4$ and $Z_0 \ge -\Gamma\sqrt{d}/4$ imply the first inequality; the usual Gaussian maximum bound implies the second. Their total failure probability is at most $\delta/4$ when $d \ge K\log((M+1)/\delta)/\Gamma^2$. The intersection thus has probability at least $1-\delta$. It is built from fixed coefficient directions before considering the actual updates. All subsequent arguments are deterministic on this intersection.

Finite attention gates and destination writes. Let $e_c = \|\widetilde{q}_c - q_{a_c}\|$ and assume $e_c \le \rho/32$. Every source score changes by at most $2e_c$, leaving threshold margin at least $\alpha_\star/16$. Thus (32) has active gates in $[1-h, 1]$ and inactive gates in $[0, h]$, where $h = \exp(-\beta\alpha_\star/16)$. Let L bound the number of edge tokens per call and let $r_{\max} = \max_c r_c$. With $h \le 1/4$, the inactive mass of the destination head is at most Le^{-T}. Among active edges, its logits vary by at most $2Th$, so its conditional distribution has ℓ_1 distance at most $e^{2Th}-1$ from the uniform distribution. For $2Th \le 1$, this distance is at most $4Th$. Since every destination value has norm at most 2, after output scaling,

$$\left\|\widetilde{o}_c - G\frac{\mathbf{1}_{N_c}}{\sqrt{r_c}}\right\| \le 8\sqrt{r_{\max}}Th + 4\sqrt{r_{\max}}Le^{-T}. \tag{72}$$

This uses one active edge per new destination. For $0 < \zeta < 1$, choose

$$T \ge \max\{1, \log(8\sqrt{r_{\max}}L/\zeta)\}, \qquad h \le \min\{1/4, 1/(2T), \zeta/(16\sqrt{r_{\max}}T)\}. \tag{73}$$

A finite β achieves this bound on h; both scales are chosen from the deterministic counts and tolerances, independently of G. Hence the vector error in (34) is at most ζ.

Residual writing and normalization across all steps. The masked copy head is exact. The prescribed scalar branches and fixed output projection write $(\lambda_c - 1)\widetilde{q}_c + \eta_c \widetilde{o}_c$, which combines with the residual as (35). RMSNorm is used with fixed gain $1/\sqrt{d}$ on the d-dimensional state channel and zero normalization offset, so it returns the unit direction. For cumulative, write $m_c = |R_c|$, $b_c = \mathbf{1}_{N_c}/\sqrt{r_c}$. Disjointness gives the exact coefficient identity

$$a_{R_{c+1}} = \frac{\sqrt{m_c}\, a_{R_c} + \sqrt{r_c}\, b_c}{\sqrt{m_c + r_c}}. \tag{74}$$

Compare the actual pre-normalization state $\widetilde{z}_c = \sqrt{m_c}\widetilde{q}_c + \sqrt{r_c}\widetilde{o}_c$ with $y_c = \sqrt{m_c + r_c}\, G a_{R_{c+1}}$. By (70) and (72),

$$\frac{\|\widetilde{z}_c - y_c\|}{\sqrt{m_c + r_c}} \le e_c + \epsilon + \zeta, \qquad \frac{\|y_c\|}{\sqrt{m_c + r_c}} \ge 1 - \epsilon. \tag{75}$$

For nonzero x, y, the normalization inequality $\|x/\|x\| - y/\|y\|\| \le 2\|x - y\|/\|y\|$ therefore yields

$$e_{c+1} \le 4e_c + 4\epsilon + 4\zeta. \tag{76}$$

The small-error choices below ensure $\widetilde{z}_c \neq 0$. For frontier, $b_c = a_{F_{c+1}}$ and the old branch is discarded, giving $e_{c+1} \le 4\zeta$, which also satisfies (76). Initialize $\widetilde{q}_0 = q_{a_0}$. With $H_D = 4(4^D - 1)/3$, induction gives $e_c \le H_D(\epsilon + \zeta)$. Choose $\epsilon, \zeta \le \rho/(128 H_D)$. Then $e_c \le \rho/64$ at every step, which closes the gate-margin induction. Absorbing the absolute constant in ϵ into A_D proves (21).

Final attention and sufficient width. The perturbation of any final score is at most $2e_D$ on the key-norm event. Thus the ideal gap $\Gamma/4$ in (71) remains strictly positive: its reduction is at most $4e_D \le \Gamma/16$. The actual final head selects the useful key. This is a pathwise transfer, not a Gaussian calculation conditional on $\widetilde{q}_D$. All concentration and induction requirements hold under

$$d \ge K_D \max\left\{ \frac{\log(Dn/\delta)}{\alpha_\star^2}, \frac{\log((M+1)/\delta)}{\Gamma^2}, \frac{\log(16(D+1)/\delta)}{\rho^2} \right\}. \tag{77}$$

For fixed D, δ and $M \ge 2$, the last term and the key-norm bound are absorbed by a constant multiple of (17). This proves the actual-computation sufficient-width claim. The constant tracks the accumulation factor H_D; the theorem makes its asymptotic comparison at fixed D. □

Proof of Theorem 5.1. Write $L := \log(Dn/\delta)$ and $B := 2\log M$, and interpret a zero alignment or reusable coefficient as infinite width. For every $c < D$ and every $a_c \in \mathcal{A}_{R_c}$,

$$\left(\min_{v \in R_c} a_{c,v} \right)^2 \le \frac{1}{|R_c|} \sum_{v \in R_c} a_{c,v}^2 = \frac{1}{|R_c|}, \tag{78}$$

with equality only at $a_c = a_{R_c}$. Hence the routing part of $\mathfrak{D}(\boldsymbol{a}_R; \mathcal{U})$ is no larger than that of any $\mathfrak{D}(\boldsymbol{a}; \mathcal{U})$.

For the readout part, Theorem 5.3 gives $\Gamma(a_{R_D}; \mathcal{U}) \ge 1/C$. Cauchy–Schwarz gives $\Gamma(a_D; \mathcal{U}) \le 1$ for every unit a_D, so

$$\frac{B}{\Gamma(a_{R_D}; \mathcal{U})^2} \le C^2 B \le C^2 \frac{B}{\Gamma(a_D; \mathcal{U})^2}. \tag{79}$$

Combining the routing inequality and (79), and using $C \ge 1$, yields

$$\mathfrak{D}(\boldsymbol{a}_R; \mathcal{U}) \le C^2 \mathfrak{D}(\boldsymbol{a}; \mathcal{U}) \tag{80}$$

for every weighted sequence $\boldsymbol{a}$. Taking the infimum proves (23).

The factor is sharp. Suppose $s := |R_D|/C^2$ is an integer, choose $S \subseteq R_D$ with $|S| = s$, and let $u = \mathbf{1}_S/\sqrt{s}$ and $\mathcal{U} = \{u\}$. This class is C-nonlocalized. Cumulative has terminal alignment $1/C$, whereas the weighted terminal state $a_D = u$ has alignment one. Keep all earlier states cumulative. Whenever the readout scale B is at least the cumulative routing scale $L \max_{c<D} |R_c|$, the two objectives are respectively $C^2 B$ and B, so equality holds in (23). Such instances are compatible with the common-bank constraint: take $D = 1$, $|R_0| = 1$, and enough unused vertices that M is proportional to n. For nonintegral m/C^2, cap $\lfloor m/C^2 \rfloor$ coordinates and use one residual coordinate; as $m \to \infty$, the ratio tends to C^2.

It remains to prove exact robust optimality. Put $m := |R_D|$ and let $u_\star$ minimize $\|u\|_1$ over the nonempty compact set $\mathcal{U}_C(R_D)$. This full class is closed under every coordinate permutation P of R_D. Hence, for any $a_D \in \mathcal{A}_{R_D}$,

$$\begin{aligned}\Gamma(a_D;\mathcal{U}_C(R_D)) &\le \frac{1}{m!}\sum_P \langle a_D, Pu_\star\rangle = \frac{\|a_D\|_1 \|u_\star\|_1}{m} \\ &\le \frac{\|u_\star\|_1}{\sqrt{m}} = \Gamma(a_{R_D};\mathcal{U}_C(R_D)). \end{aligned} \tag{81}$$

The second inequality uses $\|a_D\|_1 \le \sqrt{m}\,\|a_D\|_2 = \sqrt{m}$; the final equality follows because $\langle a_{R_D}, u\rangle = \|u\|_1/\sqrt{m}$. Thus a_{R_D} minimizes the robust readout term. Equation (78) shows that every earlier a_{R_c} minimizes its routing term. The cumulative sequence therefore minimizes their maximum, proving (24).

Finally, restricting the comparison to any family containing $\boldsymbol{a}_R$ can only increase its infimum, so the same C^2 bound continues to hold. Since the full-family minimum is attained by $\boldsymbol{a}_R$, its exact minimax statement also survives this restriction. The actual cumulative construction has sufficient width $K_{D,\delta}\mathfrak{D}(\boldsymbol{a}_R;\mathcal{U})$ by Theorem 4.1. Combining these two inequalities gives the implementation guarantee stated in the remark. The precise factor C^2 concerns the defined objective $\mathfrak{D}$; the construction's stability constant is accounted for separately. □

C Exact Reduction for an Arbitrary Fixed Bank

The orthogonal bank makes the main dimension law transparent. The following identity shows that the same signal-versus-competition structure persists for an arbitrary fixed correlated bank.

Theorem C.1 (Arbitrary-bank Gaussian reduction)**.** *Fix a unit state a and deterministic candidate directions $u_0, u_1, \dots, u_M$, independent of G. For each $j \ge 1$, let*

$$h_j := u_0 - u_j, \qquad \gamma_j := \langle a, h_j\rangle > 0, \qquad r_j := h_j - \gamma_j a \in a^\perp. \tag{82}$$

Let $g \sim N(0, I_n)$ and $C_d \sim \chi_d$ be independent. Then

$$\mathbb{P}(\widehat{j}(q_a) = 0) = \mathbb{P}\left(\max_{1\le j\le M}\langle g, -r_j/\gamma_j\rangle < C_d\right). \tag{83}$$

Proof. The correct candidate beats candidate j exactly when

$$\begin{aligned} 0 &< \langle Ga, G(u_0 - u_j)\rangle \\ &= \gamma_j \|Ga\|^2 + \langle Ga, Gr_j\rangle. \end{aligned} \tag{84}$$

Since $\gamma_j > 0$, simultaneous success is equivalent to

$$\max_{j\le M}\langle Ga, G(-r_j/\gamma_j)\rangle < \|Ga\|^2. \tag{85}$$

Every r_j is orthogonal to a. Joint Gaussianity therefore makes Ga independent of the family $\{G(-r_j/\gamma_j)\}_{j\le M}$. Conditional on $Y := Ga$, the process on the left of (85) has the same law as

$$\frac{\|Y\|}{\sqrt{d}}\left\{\langle g, -r_j/\gamma_j\rangle\right\}_{j\le M}. \tag{86}$$

Canceling $\|Y\| > 0$ and using $\sqrt{d}\,\|Y\| \sim \chi_d$ proves (83). □

Theorem C.1 requires no semantic choice of query and no orthogonality. Correlation changes only the maximum of the fixed Gaussian family. In the isotropic model, that maximum is $(1/\mu)\max_{j\le M} Z_j$ up to a common centered Gaussian term, recovering Theorem A.3. More generally, the critical dimension is the squared typical scale of the Gaussian maximum in (83). The frontier–cumulative comparison remains a comparison under the same bank; the simplified $\log M/\mu^2$ form is obtained when the competition geometry is common to both states.

D Experimental Details and Additional Results

D.1 Dataset and query construction

We use the public graph-4 ProsQA splits released with Zhu et al. (2025b). Table 1 records the full split statistics. The paired width sweep uses every training and test trace and a fixed 120-trace validation subset.

Table 1: ProsQA graph-4 statistics. Graph sizes and solution lengths are averaged over traces.

Split	# traces	$\lvert V\rvert$	$\lvert E\rvert$	solution length
Train	14,785	22.8	36.5	3.5
Validation	257	22.7	36.3	3.5
Test	419	22.7	36.0	3.5

For the lightweight two-block sweep, queries are sampled in coefficient space before the Gaussian embedding. Each query contains the complete terminal frontier but remains strictly partial, uses between 75% and 85.7% of the reached history, and has within-support coefficient ratio at most $\rho = 1.2$. These choices imply the certified cap $C = 1.386$; the largest realized caps are smaller on every split. The same query, Gaussian draw, useful candidate, and decoy bank are reused for frontier and cumulative.

Table 2: Audit of the fixed nonlocalized query bank. "Queries" counts coefficient-space queries before repeated Gaussian test draws.

Split	Used traces	Queries	Support fraction	Realized $C_{\max}$
Train	14,785	118,280	$[0.750, 0.857]$	1.319
Validation	120	480	$[0.750, 0.857]$	1.271
Test	419	3,352	$[0.750, 0.857]$	1.287

D.2 Lightweight two-block width sweep

The learned construction is the two-attention router in Theorem 4.1. Its source query and key projections are initialized as $I + 0.08Z/\sqrt{d}$; the membership inverse temperature and destination temperature are initialized to 8 and 10 and trained jointly. Frontier and cumulative share all router parameters and optimizer updates. The best epoch is selected by the sum of their validation state cosines.

Table 3: Configuration of the lightweight two-block paired experiment.

Component	Configuration
Widths	$16, 24, 32, 40, 48, 56, 64, 80, 96, 112, 128$
Training seeds / embedding seeds	0:4 / 5100:5104
Query-bank size per trace (train/val/test)	$8/4/8$
Decoys (train/val/test)	$256/1{,}024/4{,}096$
Gaussian draws (val/test)	$1/4$
Epochs and optimizer	4; AdamW, learning rate 3×10^{-3}
Regularization	weight decay 10^{-4}; gradient clip 1.0
Gradient accumulation	16 traces
Loss weights (gate/state/query)	$1/2/0.1$
Query seed and certified cap	1701; $C = 1.385640646$
Primary endpoint	linearly interpolated $d_{0.9}$ at test success 0.9

At test time, success is top-1 retrieval of the fixed useful candidate against 4,096 decoys. Each training seed is evaluated on 419 traces, eight queries per trace, and four Gaussian draws. Statistical intervals are therefore computed across the five independent training seeds, rather than by treating the within-seed evaluations as independent replications. During training, each trace uses one query

per epoch, cycling deterministically through the pre-generated eight-query bank; hence four epochs expose four query IDs per trace.

D.3 Complete fixed-query results

Table 4 gives the complete width sweep underlying Figure 3. Intervals are two-sided 95% Student-t intervals across five training seeds. "Ideal" supplies the analytic state to the same decoder and is not an additional trained model.

Table 4: Held-out top-1 accuracy at every tested width.

d	Learned frontier	Learned cumulative	Ideal frontier	Ideal cumulative
16	0.048 ± 0.001	0.254 ± 0.012	0.052 ± 0.001	0.442 ± 0.003
24	0.114 ± 0.004	0.574 ± 0.008	0.120 ± 0.005	0.790 ± 0.004
32	0.210 ± 0.007	0.793 ± 0.015	0.216 ± 0.007	0.947 ± 0.005
40	0.322 ± 0.005	0.899 ± 0.010	0.328 ± 0.006	0.990 ± 0.002
48	0.430 ± 0.008	0.946 ± 0.004	0.436 ± 0.008	0.9987 ± 0.0002
56	0.533 ± 0.010	0.972 ± 0.005	0.540 ± 0.009	0.9999 ± 0.0002
64	0.622 ± 0.008	0.984 ± 0.003	0.630 ± 0.008	1.0000 ± 0.0001
80	0.758 ± 0.010	0.992 ± 0.002	0.767 ± 0.010	1.000 ± 0.000
96	0.847 ± 0.005	0.997 ± 0.001	0.855 ± 0.006	1.000 ± 0.000
112	0.905 ± 0.006	0.998 ± 0.001	0.913 ± 0.005	1.000 ± 0.000
128	0.941 ± 0.005	0.998 ± 0.001	0.949 ± 0.003	1.000 ± 0.000

The threshold is stable at the individual-seed level (Table 5). Linear interpolation gives a learned cumulative/frontier ratio of 0.365 ± 0.008 (mean and 95% interval); the fixed-alignment prediction is $(\mu_F/\mu_R)^2 = (0.5011/0.8893)^2 = 0.318$. The analytic states first cross the tested grid at widths 112 and 32 for frontier and cumulative, respectively. At $d = 40$, learned-state cosines are 0.966 ± 0.006 for frontier and 0.933 ± 0.006 for cumulative. Thus the cumulative accuracy advantage cannot be explained by a higher state cosine.

Table 5: Interpolated empirical critical dimension $d_{0.9}$ for every lightweight-model training seed.

Seed	Frontier	Cumulative	C/F
0	108.9	39.2	0.360
1	112.1	39.9	0.356
2	110.3	40.4	0.366
3	111.7	41.6	0.372
4	110.1	40.6	0.369
Mean	110.6	40.3	0.365
95% interval	± 1.6	± 1.1	± 0.008

thm:uniform-separation

D.4 GPT-2 replication: configuration and complete results

The GPT-2 replication uses the same ProsQA graph-4 traces and the same paired comparison, but replaces the lightweight router with a two-layer causal GPT-2 Transformer. Each edge is represented by causal source and destination tokens followed by one continuous-thought token. The two states share the model, optimizer updates, coefficient-space query, Gaussian codebook, useful candidate, and decoy bank; only the frontier versus cumulative residual update differs. Queries are generated before the Gaussian map, use the full reached support ($\alpha = 1$), and draw positive coefficients with imbalance at most $\rho = 1.1$.

Each seed–width model is evaluated on 120 held-out traces, four fixed queries per trace, and two Gaussian draws, with the useful candidate ranked against 4,096 fixed decoys. Table 7 reports all tested widths; intervals are two-sided 95% Student-t intervals over the five model seeds.

The seed-level interpolants in Table 8 are the raw values summarized in the main-text figure. Cumulative has a lower critical dimension for every paired seed; the mean ratio is 0.521 ± 0.067.

Table 6: Complete configuration of the paired GPT-2 width sweep.

Component	Configuration
Backbone	causal GPT-2; 2 layers, 8 heads, FFN width $4d$
Transformer details	GELU, context 512, dropout 0.1, residuals and LayerNorm
Tested widths / maximum codebook width	$32, 64, 96, 128, 192, 256, 384$ / 768
Model seeds / embedding seeds	0:4 / 6100:6104
Used traces (train/validation/test)	$800/120/120$
Queries per trace (train/validation/test)	$8/4/4$
Query construction	full reached support; $\rho = 1.1$; seed 2701
Decoys (train/validation/test)	$256/1{,}024/4{,}096$
Gaussian draws (validation/test)	$1/2$
Optimizer	AdamW; learning rate 10^{-4}; weight decay 10^{-2}
Optimization	clip 1.0; accumulation 16; at most 40 epochs
Loss weights (routing/state/query)	$1/10/0.1$
Stopping rule	after epoch 6: routing $\geq .98$ and both cosines $\geq .80$
Stopping patience / checkpoint selection	2 epochs / best summed validation cosine
Decoder and endpoint	shared top-1 decoder; interpolated $d_{0.9}$

Table 7: Complete GPT-2 held-out top-1 accuracy sweep.

d	Frontier	Cumulative
32	0.468 ± 0.038	0.898 ± 0.024
64	0.900 ± 0.026	0.998 ± 0.003
96	0.978 ± 0.009	1.000 ± 0.000
128	0.993 ± 0.004	1.000 ± 0.000
192	0.998 ± 0.002	1.000 ± 0.000
256	1.000 ± 0.001	1.000 ± 0.000
384	1.000 ± 0.000	1.000 ± 0.000

D.5 Weighted routing protocol and raw codebook results

At $d = 192$, the identity-regularized router is trained for four epochs on 600 traces and six profiles with 64 Gaussian route decoys, AdamW ($\mathrm{lr} = 3 \times 10^{-3}$, weight decay 10^{-4}), clip 1.0, accumulation 16, initialization noise 0.02, and identity penalty 10. We select by mean route success minus 0.02 projection error on 80 validation traces, 256 Gaussian decoys, and one draw. The frozen checkpoint is evaluated without retraining on five codebooks, 120 test traces, two draws, all unreached graph keys, and 512 additional Gaussian decoys per route step.

The corresponding mean minimum coefficient, normalized by the uniform value, is 1.000, 0.814, and 0.775. The monotone association between this weakest coefficient, joint success, and false positives is the token-level pattern predicted by the $a_{\min}$ mechanism in Theorem 5.1.

Table 8: GPT-2 interpolated critical dimension $d_{0.9}$ for every model seed.

Seed	Frontier	Cumulative	C/F
0	63.2	32.0	0.507
1	68.6	33.0	0.481
2	73.5	36.2	0.493
3	62.6	32.0	0.511
4	63.0	38.8	0.616
Mean	66.2	34.4	0.521
95% interval	± 5.9	± 3.8	± 0.067

Table 9: Raw reusable-token routing results at $d = 192$. Each cell reports mean joint route-step success (fraction) / per-inactive-key false-positive rate (%). A route step succeeds iff every currently reached node is accepted and every unreached graph node and all 512 Gaussian decoys are rejected.

Codebook seed	Uniform cumulative	Recency-heavy	Random mild
4903	$0.896/0.032$	$0.576/0.572$	$0.533/1.318$
5004	$0.904/0.028$	$0.594/0.585$	$0.541/1.282$
5105	$0.907/0.028$	$0.593/0.557$	$0.536/1.312$
5206	$0.913/0.026$	$0.587/0.564$	$0.541/1.336$
5307	$0.909/0.029$	$0.578/0.570$	$0.534/1.280$
Mean	$0.906/0.029$	$0.586/0.570$	$0.537/1.306$

D.6 Software and computing resources

We use Python 3.10, PyTorch 2.5.1, CUDA 12.4, NumPy 2.1.3, and Transformers 4.46.2 on NVIDIA A100 GPUs. The lightweight sweep contains 55 trained models and records 37.6 summed GPU-hours; the GPT-2 replication contains 35 single-GPU training runs. The four-width weighted run records 0.17 GPU-hours and its token analysis reuses the $d = 192$ checkpoint.