TRANSFORMERS PROVABLY SOLVE PARITY BY INTERNALIZING CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) improves the performance of transformers on complex reasoning tasks. However, generating explicit reasoning traces incurs substantial inference overhead. Implicit CoT (ICoT) aims to internalize these reasoning steps by gradually removing intermediate CoT tokens during training, yet the learning dynamics of this process remain poorly understood. We provide a theoretical analysis of ICoT for parity learning in a one-layer transformer with multi-head softmax attention. We prove that removing one intermediate token at a time enables ordinary gradient descent (GD) to solve parity with polynomially many samples and updates under suitable initialization. Our analysis characterizes a race among attention heads in which each newly missing parity bit recruits a new head while previously learned assignments are preserved. In contrast, we show that direct training from final-answer supervision alone leaves GD with high prediction error. Numerical experiments support our theoretical findings. Together, these results provide a mechanistic explanation for how a curriculum with intermediate supervision can enable efficient learning without explicit CoT at inference time.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning improves the ability of large language models (LLMs) to solve challenging tasks by generating explicit intermediate steps (Wei et al., 2022; Kojima et al., 2022). However, generating these additional tokens sequentially increases inference latency and computational cost. Latent reasoning methods instead seek to perform intermediate computation within the model's hidden representations. In particular, Deng et al. (2024) develop stepwise internalization for implicit chain-of-thought (ICoT), progressively removing intermediate CoT tokens during training until the model learns to predict the final answer directly. This curriculum seeks to preserve the benefits of intermediate supervision while eliminating explicit reasoning traces at inference time. However, a theoretical understanding of how this curriculum enables internalization remains limited.

Parity learning offers a natural testbed for this question. Kim & Suzuki (2025) show that a one-layer transformer can efficiently learn parity with CoT supervision, while direct training on final answers alone is provably harder. Recent work has attempted to analyze the training dynamics of ICoT using parity learning. Huang et al. (2026) prove internalization in a gated multi-layer transformer using a curriculum that replaces groups of CoT tokens with zero padding and rounds attention weights after each gradient update. Tsilivis et al. (2026) analyze one-token deletion in a one-layer linear-attention model, with parameter rounding at stage ends. However, these guarantees rely on additional architectural and training mechanisms, such as fixed gates, specialized attention masks, and parameter rounding, which limit how closely they reflect standard ICoT training.

We study the coupled training dynamics of multi-head softmax attention during stepwise CoT internalization. To model the deletion curriculum, we replace the binary-tree parity decomposition of Kim & Suzuki (2025) with a sequential prefix-product chain (Figure 1) and introduce a BOS token that serves as an attention sink for heads not currently needed. We then characterize how ordinary gradient descent (GD) progressively recruits attention heads as intermediate tokens are removed, while preserving previously learned assignments. This process ultimately yields a model that computes parity directly from the input, without requiring CoT tokens at inference.

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 contributions are summarized as follows:

- We prove that stepwise internalization enables ordinary GD to learn parity in a one-layer multi-head softmax attention transformer with polynomially many samples and updates under suitable initialization without CoT at inference (Theorem 1). In contrast, direct final-answer training under population GD retains high prediction error (Theorem 2).
- We characterize the coupled attention dynamics underlying internalization: each stage recruits exactly one idle head for the newly required parity bit while preserving earlier assignments, even as previously recruited heads continue to train (Section 5). We further analyze how initialization and curriculum design affect internalization (Sections 4.3 and 4.4).
- We conduct numerical experiments that demonstrate successful internalization and support the predicted training dynamics (Section 6).

2 Related work

CoT and latent reasoning. CoT prompting makes intermediate computations explicit (Wei et al., 2022). Expressivity analyses show that these tokens increase the computational power of bounded-depth transformers (Feng et al., 2023; Merrill & Sabharwal, 2024). Beyond natural-language reasoning traces, pause and filler tokens provide additional computation positions without requiring meaningful intermediate text (Goyal et al., 2024; Pfau et al., 2024). Coconut feeds back continuous hidden states (Hao et al., 2025), while CODI transfers explicit CoT reasoning into continuous latent states through self-distillation (Shen et al., 2025). These methods change how intermediate computation is represented while retaining additional positions or latent steps at inference.

Training dynamics. The training dynamics of transformers have been studied from several complementary perspectives (Bietti et al., 2023; Boix-Adserà et al., 2023; Zhang et al., 2024; Wen et al., 2025). One line of work explains how attention patterns emerge during training, including the selection of relevant tokens (Tian et al., 2023; Wang et al., 2024). Another examines how transformers learn structured computations, such as recovering latent causal structure (Nichani et al., 2024) and composing permutations (Wang et al., 2025). The closest work to our setting is Kim & Suzuki (2025), which establishes that CoT enables efficient parity learning while direct final-answer training is provably harder. We extend this separation to ICoT, showing that its curriculum preserves efficient learning while removing the need for CoT at inference.

CoT internalization. Deng et al. (2024) demonstrate that models can internalize CoT through progressive left-to-right token deletion. The closest theoretical results study this transition on parity. Huang et al. (2026) introduce Log-ICoT, which replaces successive levels of a parity tree with zero padding. Their guarantee uses a gated multi-layer transformer and rounded gradient updates, and retains padded positions at inference. Tsilivis et al. (2026) analyze one-token prefix deletion in a one-layer linear-attention model, updating only the first output row after the initial stage and rounding parameters at stage ends; their experiments also succeed without rounding. Our guarantee covers one-token deletion in a multi-head softmax model under ordinary GD, without rounding or freezing trainable parameters. Table 1 summarizes these differences.

Table 1: Comparison of existing theoretical analyses of CoT and ICoT for parity learning.

	Kim & Suzuki (2025)	Huang et al. (2026)	Tsilivis et al. (2026)	This work
Layers	1	$\log_2 k$	1	1
Attention	single-head softmax	single-head softmax	single-head linear	multi-head softmax
Link function	two-bit parity	two-bit parity	piecewise-linear parity	product (XOR)
Curriculum	—	zero-padding	one-token deletion	one-token deletion
Training update	approx. GD	approx. GD + rounding	GD	GD
Between stages	—	readout switch	integer rounding	—
Inference	explicit CoT	one pass	one pass	one pass

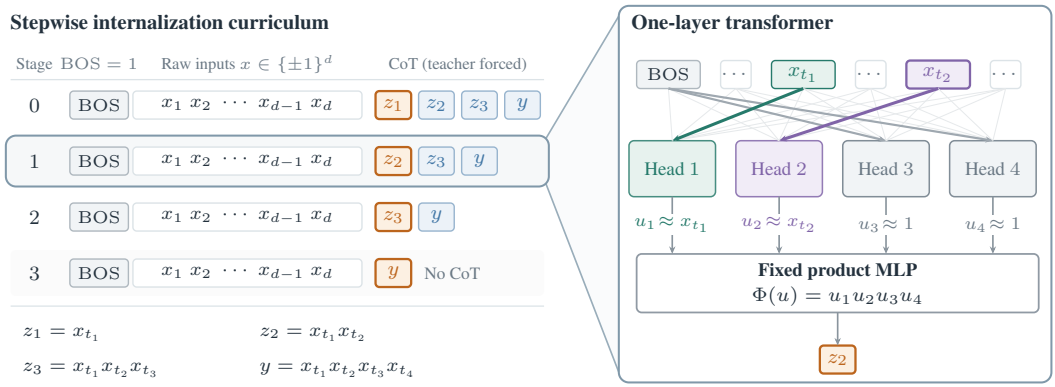


Figure 1: Stepwise internalization ($k = H = 4$), with the row-0 model after stage 1 at right. Purple marks new recruitment, green retained assignments, and orange targets; heads are relabeled by recruitment order.

3 Preliminaries

Notation. For a positive integer N, write $[N] = \{1, \ldots, N\}$. We use $\mathbf{1}_n$ for the all-ones vector in $\mathbb{R}^n$ and $\odot$ for entrywise multiplication. The norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the Euclidean and maximum norms, respectively.

3.1 Task settings

We study k-parity learning in input dimension $d \geq k \geq 2$. Fix a tuple $t = (t_1, \ldots, t_k)$ of distinct indices in $[d]$, unknown to the learner and fixed across training samples. We call t the *secret tuple* (or *secret*) and its entries the *secret keys*, identifying the k relevant input bits. Given $x \sim \mathrm{Unif}(\{\pm 1\}^d)$, the target is $y = \prod_{i=1}^k x_{t_i}$.

Task decomposition. We supply CoT supervision through successive prefix products:

$$z_1 = x_{t_1}, \qquad z_m = z_{m-1} x_{t_m} \quad (2 \leq m \leq k), \qquad z_k = y. \tag{1}$$

The supervised sequence is $(z_1, \ldots, z_{k-1}, y)$, where $z_1, \ldots, z_{k-1}$ are the intermediate CoT tokens. The one-bit prefix z_1 supplies the initial learning target.

BOS token. We add a beginning-of-sequence (BOS) token with fixed value 1, available at every prediction position and retained throughout the curriculum. It acts as an *attention sink*, absorbing attention when a head is not needed for the current computation; a head concentrated on BOS contributes approximately 1 to the product readout. This role is motivated by attention-sink behavior in pretrained transformers, where initial tokens attract large attention weights despite limited semantic relevance (Xiao et al., 2024; Sun et al., 2024; Guo et al., 2024; Gu et al., 2025).

Curriculum. At stage $s \in \{0, \ldots, k-1\}$, the first s intermediate tokens have been removed, leaving $(z_{s+1}, \ldots, z_k)$ as the prediction sequence. We index its positions by rows $j \in \{0, \ldots, k-1-s\}$. Under teacher forcing, row j predicts z_{s+j+1} from the raw input and the preceding ground-truth CoT tokens:

$$\mathcal{K}_{s,j} = [\mathrm{BOS}, x_1, \ldots, x_d, z_{s+1}, \ldots, z_{s+j}], \qquad L_j = 1 + d + j. \tag{2}$$

Here L_j is the number of available keys. Row 0 always sees the same $L = d + 1$ keys $[\mathrm{BOS}, x_1, \ldots, x_d]$. Its target grows from z_1 to z_k, adding one factor per stage; at the final stage, it predicts y without intermediate CoT tokens.

3.2 Model

Softmax attention layer. We use a one-layer attention-table model with $H \geq 1$ softmax heads followed by a fixed feedforward link. Each absolute prediction row $j \in \{0, \ldots, k-1\}$ and head

$h \in [H]$ has a trainable vector $w^{(j,h)} \in \mathbb{R}^{d+j}$ of non-sink logits and a fixed sink score b_h. Let W collect all trainable logits, with row blocks $w^{(j,\cdot)} = (w^{(j,h)})_{h=1}^{H}$. Raw inputs occupy fixed slots; deletion shifts the surviving CoT values while retaining the parameters at their original row and key slots. For the key list in eq. (2), the attention probabilities are

$$p_{\mathrm{BOS}}^{(j,h)} = \frac{e^{b_h}}{Z_{j,h}}, \qquad p_{\kappa}^{(j,h)} = \frac{e^{w_{\kappa}^{(j,h)}}}{Z_{j,h}}, \qquad Z_{j,h} = e^{b_h} + \sum_{\kappa \neq \mathrm{BOS}} e^{w_{\kappa}^{(j,h)}}. \tag{3}$$

For analytical convenience, we use the BOS logit b_h as a fixed softmax reference. Since softmax depends only on logit differences, this does not restrict the representable attention distributions; the BOS attention probability can still change during training.

Here κ ranges over the non-sink keys in $\mathcal{K}_{s,j}$. A key subscript denotes the slot currently occupied by that key; raw key x_b is indexed by $b \in [d]$.

For a batch $x^1, \ldots, x^n$, let x_b^i and z_m^i denote input coordinate b and prefix product m of sample i. Collect the key values into $v_{\mathrm{BOS}} = \mathbf{1}_n$, $v_b = (x_b^i)_{i=1}^n$, and $v_{z_m} = (z_m^i)_{i=1}^n$. For $S \subseteq [d]$, write $v_S = \bigodot_{b \in S} v_b$, with $v_\varnothing = \mathbf{1}_n$; in particular, $v_{z_m} = v_{\{t_1,\ldots,t_m\}}$. The batch output of head h at row j is

$$u^{(j,h)} = \sum_{\kappa \in \mathcal{K}_{s,j}} p_{\kappa}^{(j,h)} v_{\kappa}.$$

Feedforward layer. Following Kim & Suzuki (2025), we model the feedforward (MLP) sublayer through a fixed link function. Here $\Phi : [-1,1]^H \to [-1,1]$ is the product map $\Phi(a_1, \ldots, a_H) = \prod_{h=1}^{H} a_h$. Applied to the H head outputs separately at each row and for each sample, it gives

$$\hat{z}_i^{(j)} = \Phi\big(u_i^{(j,1)}, \ldots, u_i^{(j,H)}\big), \qquad \hat{z}^{(j)} = \bigodot_{h=1}^{H} u^{(j,h)}. \tag{4}$$

This fixed product readout makes a head concentrated on a raw key contribute that bit, while a head concentrated on the sink contributes the multiplicative identity.

For example, an MLP with square activation $\sigma(t) = t^2$ can compute Φ by composing two-neuron multiplication blocks based on the identity

$$ab = \tfrac{1}{4}\big[\sigma(a+b) - \sigma(a-b)\big].$$

Arranging these blocks in a balanced binary tree, with constant inputs 1 added as needed, gives an MLP with $O(H)$ neurons and $O(\log H)$ hidden layers.

Training objective. For the theoretical analysis, at each stage we draw n fresh independent uniform inputs and hold the batch fixed while applying full-batch Euclidean GD to the teacher-forced squared loss

$$\mathcal{L}^{(s)}(W) = \sum_{j=0}^{k-1-s} \ell_j^{(s)}(w^{(j,\cdot)}), \qquad \ell_j^{(s)} = \frac{1}{2n} \left\| \hat{z}^{(j)} - v_{z_{s+j+1}} \right\|_2^2. \tag{5}$$

All active non-sink logits are updated with step size η. The row blocks are disjoint, and teacher-forced key values are parameter-independent, so the rows evolve independently (Lemma 2). This lets the analysis follow row 0 of Algorithm 1 under the full-loss updates.

For this row, we abbreviate $w = w^{(0,\cdot)}$, $w_h = w^{(0,h)}$, $p_{h\kappa} = p_{\kappa}^{(h)} = p_{\kappa}^{(0,h)}$, and $u^{(h)} = u^{(0,h)}$. The raw-key logits are w_{hb} for $b \in [d]$, and $\ell_s(w) = \ell_0^{(s)}(w)$. Write $u_h(x)$ for a single-input head output and $\hat{y}_W(x) = \prod_h u_h(x) = \hat{z}^{(0)}(x; W)$ for the predictor; its batch output is $\hat{y} = (\hat{y}_W(x^i))_{i=1}^n = \hat{z}^{(0)}$.

Stage transitions. The stopping rule first identifies a newly recruited head when exactly one formerly idle head places more than half its attention on a raw key. Training then continues until the assignment is sufficiently sharp. At a nonfinal stage, the rule also preserves the remaining heads' ability to learn a future key. The exact inequalities are given in Definition 1. The final total-error threshold is $d^{-(c+2)}$, where $c \geq 1$ is the accuracy exponent in Theorem 1.

Algorithm 1 Stepwise internalization

Require: Data with CoT supervision, batch size n, step size η, per-stage cap $T_{\max}$

Initialize W once using eq. (6); start with no recruited heads.
for $s = 0, \dots, k-1$ **do**
  Draw n fresh inputs; form $\mathcal{L}^{(s)}$ from the remaining CoT sequence (eq. (5)).
  for $q = 1, \dots, T_{\max}$ **do**
    $W \leftarrow W - \eta \nabla_W \mathcal{L}^{(s)}(W)$ ▷ Update all active rows.
    if the attention stopping rule in Definition 1 holds **then**
      break ▷ The new assignment is sufficiently sharp.
  if the stopping rule has not been met **then**
    return FAIL
  If $s < k-1$, delete z_{s+1}; retain all parameters and assignments.
return the row-0 predictor $x \mapsto \hat{y}_W(x)$, which uses no CoT.

4 Main results

4.1 Internalization

Choose positive scales $\Lambda_1, \dots, \Lambda_H$ independently of the secret and initialize

$$w_\kappa^{(j,h)} = 0 \quad (\kappa \neq \mathrm{BOS}), \qquad b_h = \log \Lambda_h. \tag{6}$$

On row 0, this gives

$$p_{\mathrm{BOS}}^{(h),0} = \frac{\Lambda_h}{\Lambda_h + d}, \qquad p_j^{(h),0} = \frac{1}{\Lambda_h + d}, \qquad \frac{p_{\mathrm{BOS}}^{(h),0}}{p_j^{(h),0}} = \Lambda_h.$$

The scale is the initial sink-to-input odds: a larger Λ_h makes a head slower to leave the sink. Every head treats all raw keys equally. For fixed $K \geq 1$ and sufficiently large B, assume

$$d^B \leq \Lambda_h \leq d^{B+K}, \qquad |\Lambda_h - \Lambda_{h'}| \geq d^{B-K} \quad (h \neq h'). \tag{7}$$

These conditions keep the scales polynomially bounded and separated. They are invariant under relabeling the heads; neither the initialization nor the algorithm orders them.

Theorem 1 (Stepwise internalization)**.** *Fix $c \geq 1$, $K \geq 1$, and $k \leq H \leq d$. For sufficiently large fixed $B = B(c, K)$ and all sufficiently large d, initialize by eqs.* (6) *and* (7)*. There are a fixed inverse-polynomial step size and polynomial sample and iteration budgets such that, for any fixed secret tuple, Algorithm 1 succeeds with probability at least $1 - kd^{-(c+4)}$. Exactly one idle head is recruited per stage, older assignments persist, and the final predictor satisfies*

$$\sup_{x \in \{\pm 1\}^d} \left| \hat{y}_W(x) - \prod_{i=1}^{k} x_{t_i} \right| \leq d^{-c}.$$

All constants are uniform over all secret tuples and scale collections satisfying eq. (7)*.*

Training parameters are given in Specification 1 (Appendix A).

The conditions allow, for example, $\Lambda_h = d^B h^a$ with any fixed $a \geq 1$ and $K \geq \max\{1, a\}$, in any head order. They also allow independently sampled scales.

Corollary 1 (Random head scales)**.** *Fix $c \geq 1$, take $K \geq c + 7$ and B as in Theorem 1, and independently sample $\Lambda_h = d^B U_h$ with $U_h \sim \mathrm{Unif}[1, 2]$. The conclusion of Theorem 1 holds with probability at least $1 - kd^{-(c+4)} - d^{-(c+5)}$ over initialization and samples.*

Only the initial sink scales are randomized; the raw-key logits remain zero as in eq. (6).

4.2 End-to-end training

For $A \subseteq [d]$, write $\chi_A(x) = \prod_{j \in A} x_j$ and define the no-CoT population risk

$$\mathcal{R}_A(W) = \frac{1}{2} \mathbb{E}_x (\hat{y}_W(x) - \chi_A(x))^2. \tag{8}$$

Population GD uses the exact gradient of this expected loss over uniform inputs; *empirical GD* uses a finite-sample average. *Population gradient flow* is the continuous-time evolution $\dot{W} = -\nabla \mathcal{R}_A(W)$.

An initialization is *key-permutation invariant* if the raw-key logits are equal within each head; the levels and sink scores may differ between heads.

Theorem 2 (End-to-end population obstruction)**.** *Fix $\alpha \in (0,1)$, $2 \le k \le d$ with $k \ge \alpha d$, $1 \le H \le d$, and any $A \subseteq [d]$ with $|A| = k$. From any key-permutation-invariant initialization, exact GD on eq.* (8)*, with any nonnegative step-size sequence, satisfies at every iterate*

$$\mathbb{E}_x(\hat{y}_{W^r}(x) - \chi_A(x))^2 \ge 1 - 2e^{-c_\alpha d}, \qquad c_\alpha = 2\alpha^2 e^{-4/\alpha}.$$

The same bound holds along population gradient flow.

The target breaks symmetry between A and its complement, but not within either set. Consequently, a head cannot select one secret key over another. The target Fourier coefficient is bounded by the probability that at most d uniform draws cover all k secret keys, which is exponentially small when $k \ge \alpha d$. The curriculum avoids this obstruction by exposing a singleton deficit at each stage.

Corollary 2 (Matched population separation)**.** *Under Theorem 1, with $k \ge \alpha d$, population GD with the curriculum succeeds deterministically in polynomial time. End-to-end population GD from the same initialization satisfies Theorem 2 at every iterate. This also holds for every random-scale draw satisfying eq.* (7)*.*

The head count in the positive theorem is also necessary for this architecture.

Proposition 1 (Head-count lower bound)**.** *The row-0 predictor has Boolean Fourier degree at most H. If $H < k$, then $\mathbb{E}_x(\hat{y}_W(x) - \chi_A(x))^2 \ge 1$ for every parameter table and every $|A| = k$.*

Proof. Each head is affine in the raw bits, so the product has degree at most H. Boolean reduction cannot increase degree. For $H < k$, orthogonality to χ_A gives squared error $1 + \mathbb{E}_x \hat{y}_W(x)^2 \ge 1$. □

4.3 Right-to-left deletion

Consider removing $z_{k-1}, \ldots, z_1$ while retaining y and the absolute parameter rows of Section 3, without copying parameters between rows.

Corollary 3 (Right-to-left population obstruction)**.** *Fix $\alpha \in (0,1)$, $3 \le k \le d$ with $k - 1 \ge \alpha d$, and $1 \le H \le d$. For the right-to-left schedule above, start from a key-permutation-invariant row-0 initialization and apply exact population GD with any nonnegative step-size sequence and arbitrary numbers of updates at each stage. At every training iterate, including after the final deletion, each row-0 head assigns equal attention to $t_2, \ldots, t_k$, and*

$$\mathbb{E}_x \left(\hat{y}_{W^r}(x) - \prod_{i=1}^{k} x_{t_i} \right)^2 \ge 1 - 2e^{-c_\alpha d}, \qquad c_\alpha = 2\alpha^2 e^{-4/\alpha}.$$

The same conclusion holds for population gradient flow.

Proof sketch. Row decoupling keeps row 0 learning z_1 until the final deletion; both z_1 and y preserve symmetry among $t_2, \ldots, t_k$, so the coverage argument of Theorem 2 applies to these $k - 1$ keys. The full proof is in Section F.1. □

4.4 Initialization

The positive theorem uses distinct heads that initially contribute approximately the multiplicative identity. The following results separate these two requirements. The first concerns permutations of heads, rather than the input-key permutations in Theorem 2.

Proposition 2 (Identical heads cannot specialize)**.** *Suppose all row-0 heads have the same fixed sink score and the same initial raw-logit vector. Exact GD with a common step size across heads preserves their equality, for empirical or population losses and across curriculum stages. At every resulting iterate and for every parity with $|A| \ge 2$,*

$$\mathbb{E}_x(\hat{y}_W(x) - \chi_A(x))^2 \ge \frac{1}{2}.$$

Head symmetry constrains the predictor to $u(x)^H$ for one affine head output u. For even H this is nonnegative; for odd H it is coordinatewise nondecreasing. Neither can approximate parity arbitrarily well. The initial logits may vary across keys, so this obstruction does not depend on raw-key symmetry.

Breaking head symmetry alone does not ensure efficient learning. When attention is spread across many input bits with little mass on BOS, each head's output is typically small. Since a head's gradient contains the outputs of all other heads, this can make learning exponentially slow even when the heads have different initial parameters.

Proposition 3 (Slow learning from random initialization)**.** *Fix $\alpha \in (0, 1]$ and $q > 0$. There are constants $\gamma, C > 0$ such that, for $\alpha d \le H \le d$ and all sufficiently large d, the following holds. Set $b_h = 0$ and initialize raw logits independently from $N(0, 1)$ or $\mathrm{Unif}[-1, 1]$. With probability at least $1 - d^{-q}$, population GD with nonnegative step sizes on the first one-bit target satisfies*

$$\mathbb{E}_x(\hat{y}_{W^r}(x) - x_{t_1})^2 \ge 1 - Ce^{-\gamma d} \quad \textit{for all } r \textit{ with } \sum_{m<r} \eta_m \le e^{\gamma d}.$$

No head reaches raw-key attention $1/2$ in this interval. For empirical GD using polynomially many independent uniform training inputs, also independent of the initialization, the same conclusions hold with an additional $e^{-\Omega(d)}$ failure probability.

The proof controls gradients throughout a neighborhood of initialization (Section E). Together, Propositions 2 and 3 motivate distinct heads whose unused state contributes approximately the multiplicative identity.

5 Proof overview

The proof of our main result, Theorem 1, proceeds by induction over the deletion stages. We first reduce the analysis to row 0 and control empirical correlations along its trajectory. At each stage, we identify a gradient advantage for the newly missing parity bit and show that head competition recruits exactly one idle head while earlier assignments survive. We then prove that sharpening restores the conditions for the next stage and that the cumulative training cost remains polynomial. Final sharpening bounds the total attention error, yielding the theorem's uniform prediction guarantee.

5.1 Gradient signal for the next parity bit

Let $A_s = \{t_1, \dots, t_{s+1}\}$ and $A_{-1} = \varnothing$. Suppose the recruited heads represent A_{s-1} and the remaining heads attend mostly to the sink. Factoring out the learned characters leaves the deficit $t = t_{s+1}$. For an idle head, define the normalized gradient $a_{hb} = p_{hb}^{-1} \partial \ell_s / \partial w_{hb}$. Then

$$a_{hb} = -1 - \mathbf{1}\{b = t\} + O(\kappa_n + \bar{\delta} + k\rho),$$

where $\bar{\delta}$ is the total off-sink mass of the idle heads and ρ bounds each recruited head's error (Lemma 3). Both target and distractor logits may initially increase, but the target has an order-one advantage. This signal identifies the next bit without any target information in the initialization.

5.2 Head specialization and assignment stability

Let $r_h = p_{ht}$ for idle heads. Neglecting the controlled distractor and old-head errors, the target masses follow

$$\dot{r}_h = (F + 1) F_{-h} r_h^2 (1 - r_h)^2, \qquad F = \prod_{h \in \mathcal{I}} (1 - 2r_h), \qquad F_{-h} = \prod_{g \in \mathcal{I} \setminus \{h\}} (1 - 2r_g). \quad (9)$$

For $G(r) = 1/[r(1 - r)]$, the head-dependent factor cancels:

$$\frac{\mathrm{d}}{\mathrm{d}t} G(r_h) = -(F + 1) F.$$

Thus all heads have the same leading reciprocal drift. A positive initial reciprocal gap keeps the leading head ahead. When that head crosses $1/2$, the sign of the target drift reverses for the followers

and returns them toward the sink. Lemma 9 proves the same conclusion for discrete GD, controlling empirical correlations, distractors, and the one-step remainders throughout the competition.

The old heads cannot simply be treated as fixed. For an old assignment with error e, its adverse reciprocal change before the new head crosses is at most

$$\frac{1}{e^{\text{cross}}} \geq \frac{1}{e^{\text{in}}} - \frac{1}{2}\{G(r^{\text{in}}) - 4\} - \text{controlled error.}$$

The stopping rule supplies $e^{\text{in}} \leq (2 - d^{-2})r^{\text{in}}$, leaving an inverse-polynomial margin in this comparison. A simultaneous stopping argument controls recruitment and the errors of previously assigned heads.

5.3 Convergence across curriculum stages

Near a correct assignment, every assigned-to-wrong logit gap is nondecreasing. If $E = \sum_h (1 - p_{h,a_h})$ is the total assignment error, then

$$\frac{c_1\eta}{HL}E^3 \leq E - E^+ \leq c_2\eta E^3$$

for universal positive constants (Lemma 5). This supplies both final accuracy and the sharpening needed before the next deletion.

The remaining issue is cumulative cost. For the analysis only, relabel the scales so that $\Lambda_1 < \cdots < \Lambda_H$, and track $X_{h\kappa} = p_{h\text{BOS}}/p_{h\kappa}$. The differences between heads stay close to their initial values, while future keys within a head remain nearly uniform. The stopping rule also preserves each surviving future reciprocal by a factor at least $1 - d^{-2}$ per stage, so an absolute lower bound does not require large ratios between adjacent scales.

Let Z_s be the leading reciprocal at entry to stage s. If $u = d - s - 1 \geq 2$ keys remain after capture, the recovery comparison gives

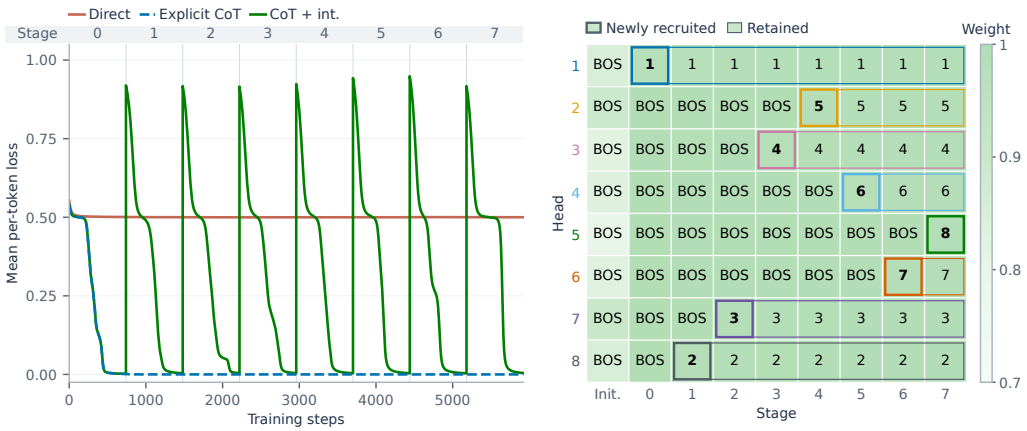
$$Z_{s+1} \leq \left(\frac{u}{u-1} + O(d^{-2})\right) Z_s + \left(\frac{2u}{u-1} + O(d^{-2})\right)(\Lambda_{s+2} - \Lambda_{s+1}) + \text{error.}$$

The multiplicative factors have product $O(d)$, and the initial gaps sum to at most Λ_H. The single possible $u = 1$ transition has an absolute-factor bound. Lemma 10 controls the cumulative errors and closes the induction for arbitrary scales satisfying eq. (7). Finally, $|\hat{y}(x) - y(x)| \leq 2E$ for every input, so the final stopping threshold implies the pointwise guarantee.

6 Experiments

We compare Direct, explicit CoT, and stepwise internalization on sparse parity with $H = k$. Direct receives only final-answer supervision; explicit CoT learns the full prefix chain; internalization progressively deletes intermediate tokens while retaining the learned parameters. Methods share their initialization, minibatches, and number of training updates. We initialize attention to concentrate on BOS, as in the theory, and use a fixed number of training updates per stage, following the token-removal schedule of Deng et al. (2024). Full details are in Section G.

Training dynamics. Explicit CoT and internalization achieve perfect final-answer accuracy, while Direct remains near chance (Table 2). Internalization achieves this accuracy with all intermediate tokens removed, showing that the curriculum can eliminate the generated reasoning chain at inference. In Figure 2, each new bit recruits a head while earlier assignments persist (Figure 3), supporting incremental internalization with Adam and trainable sinks. With $H = k+1$ or $H = k+2$, recruitment likewise leaves surplus heads on BOS (Section G.6).



(a) Average per-token loss

(b) Head assignments

Figure 2: Stepwise internalization at $d = 32$, $k = H = 8$. (a) Training restores low loss after each token deletion. (b) Each stage recruits one head while preserving earlier assignments.

Table 2: Internalization matches explicit CoT accuracy without intermediate tokens at inference, while Direct remains near chance. Final-answer test accuracy (%), mean ± SD over 10 random seeds.

Method	CoT tokens	Stages	$d = 32$			$d = 64$		
			$k = 8$	$k = 12$	$k = 16$	$k = 8$	$k = 12$	$k = 16$
Direct	0	1	50.0 ± 0.1	50.1 ± 0.2	49.9 ± 0.1	50.0 ± 0.2	50.1 ± 0.2	50.0 ± 0.1
Explicit CoT	$k - 1$	1	100.0	100.0	100.0	100.0	100.0	100.0
CoT + internalization	0	k	100.0	100.0	100.0	100.0	100.0	100.0

Head asymmetry and recruitment. At $d = 32$, $k = H = 8$, we vary initial differences between heads while retaining their BOS preference. Nearly identical heads take longer to specialize in controlled first-stage races (Section G.3). Repeating the full curriculum with the original and doubled per-stage budgets illustrates recovery: a run that remains at chance with shorter stages reaches perfect final accuracy with longer stages (Section G.4). A fixed deletion schedule can advance before a stage is learned; this timing mismatch offers a possible mechanism for the token-removal schedule sensitivity reported by Deng et al. (2024).

Random initialization. We repeat the comparison in Table 2 at $d \in \{32, 64\}$ and $k = H \in \{8, 12, 16\}$, removing only the initial BOS preference: all logits, including BOS, are drawn independently from the same Gaussian distribution. Under the unchanged training budget, internalization succeeds less consistently. At $d = 32$, $k = H = 8$, heads can form cancelling pairs and later separate as the target grows. Doubling the per-stage budget improves success from $8/10$ to $10/10$ trials (Section G.5). Thus BOS-biased initialization improves reliability, while successful internalization also extends to other patterns of head specialization.

7 Conclusion

Our multi-head softmax model provides a tractable testbed for studying how CoT supervision is internalized during training. Under suitable initialization, ordinary GD learns parity in polynomial time by recruiting heads as intermediate tokens are removed while preserving earlier assignments. The final predictor requires no CoT, and training requires no parameter rounding. For dense parity, direct population GD from the same initialization retains high error, establishing a benefit of the curriculum within the same architecture. Experiments support these dynamics and highlight sensitivity to initialization and deletion timing. This combination of provable guarantees and observable head dynamics enables a controlled study of how internalization succeeds.

Reproducibility statement

Section 3 specifies the model and training algorithm, and Section A gives the theoretical parameters and stopping rule. Complete proofs appear in Sections B to F. Section G details the synthetic data, initialization, optimization, evaluation, and additional experiments.

AI use statement

In this work, we used generative AI tools to aid or polish writing. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

References

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588, 2023. URL https://doi.org/10.52202/075280-0077.

Enric Boix-Adserà, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24519–24551, 2023. URL https://doi.org/10.52202/075280-1066.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit CoT to implicit CoT: Learning to internalize CoT step by step. arXiv:2405.14838, 2024. URL https://arxiv.org/abs/2405.14838.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70757–70798, 2023. URL https://doi.org/10.52202/075280-3100.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, pp. 27896–27923, 2024. URL https://openreview.net/forum?id=ph04CRkPdC.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, pp. 97114–97144, 2025. URL https://arxiv.org/abs/2410.10781.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. arXiv:2410.13835, 2024. URL https://arxiv.org/abs/2410.13835.

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. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Yixiao Huang, Hanlin Zhu, Zixuan Wang, Jiantao Jiao, Stuart Russell, Somayeh Sojoudi, and Song Mei. Transformers provably learn to internalize chain-of-thought. arXiv:2605.28600, 2026. URL https://arxiv.org/abs/2605.28600.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *International Conference on Learning Representations*, pp. 44642–44668, 2025. URL https://openreview.net/forum?id=n2NidsYDop.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. URL https://doi.org/10.52202/068431-1613.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008. URL https://doi.org/10.1017/CBO9781107359949.008.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, pp. 7690–7706, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38018–38070. PMLR, 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.17762.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71911–71947, 2023. URL https://doi.org/10.52202/075280-3147.

Nikolaos Tsilivis, Nirmit Joshi, Marko Medvedev, Julia Kempe, and Nati Srebro. Learning through internalization. arXiv:2606.20937, 2026. URL https://arxiv.org/abs/2606.20937.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51854–51912. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24ca.html.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D. Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5632–5711. PMLR, 2025. URL https://proceedings.mlr.press/v291/wang25a.html.

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, pp. 24824–24837, 2022. URL https://doi.org/10.52202/068431-1800.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *International Conference on Learning Representations*, pp. 100857–100901, 2025. URL https://openreview.net/forum?id=AmEgWDhmTr.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024. URL https://arxiv.org/abs/2309.17453.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL https://www.jmlr.org/papers/v25/23-1042.html.

A TRAINING PARAMETERS AND STOPPING RULE

We first specify the constants and stopping conditions used in Theorem 1. Throughout the positive proof, write

$$\Lambda_- := \min_h \Lambda_h, \qquad \Lambda_+ := \max_h \Lambda_h, \qquad g_* := \min_{h \neq h'} |\Lambda_h - \Lambda_{h'}|, \qquad \delta_* := \frac{d}{\Lambda_- + d}.$$

The largest initial off-sink mass of a row-0 head is δ_*. Equation (7) gives $\delta_* \le d^{1-B}$ and initial total error at most $H\delta_* \le d^{2-B}$. Set

$$\tau = d^{-2}, \qquad \chi = d^{11}\delta_*, \qquad R = 4(B + K + c + 30). \tag{10}$$

It suffices to take $B \ge 6K + 2c + 100$ and then d sufficiently large. The exponents are chosen for the estimates below rather than to optimize the rate.

Specification 1 (Training parameters)**.** Let $p = d^{-(c+4)}$, choose a fixed $Q \ge 8R$, and take

$$\eta = d^{-Q}, \qquad \kappa_n = \sqrt{\frac{2}{n} \log \frac{2^{d+1}}{p}} \le d^{-Q}, \qquad T_{\max} = \left\lceil d^R/\eta \right\rceil. \tag{11}$$

It suffices that $n \ge 2d^{2Q} \log(2^{d+1}/p)$. The fresh inputs drawn at each stage are independent uniform Boolean samples. All constants may depend on c, K, B, but not on the secret tuple or the particular scales satisfying eq. (7).

Unless stated otherwise, attention quantities in the proof refer to row 0. For vectors in $\mathbb{R}^n$, let $\langle v_1, \dots, v_r \rangle = \sum_{i=1}^n \prod_{a=1}^r v_{a,i}$ and $\langle v \rangle = \sum_i v_i$; products of sample vectors are entrywise. A partial assignment is a bijection $\pi : \mathcal{L} \to C$ between already recruited heads and already learned raw keys; the remaining heads form $\mathcal{I} = [H] \setminus \mathcal{L}$. Write

$$e_h = 1 - p_{h,\pi(h)} \quad (h \in \mathcal{L}), \qquad m_h = 1 - p_{h\mathrm{BOS}} \quad (h \in \mathcal{I}), \qquad \bar{\delta} = \sum_{h \in \mathcal{I}} m_h.$$

For a full assignment $a_h \in \{\mathrm{BOS}, 1, \dots, d\}$ let $E = \sum_h (1 - p_{h,a_h})$. The constant $E_0 > 0$ below is the universal vertex radius supplied by Lemma 5.

Definition 1 (Stage stopping rule)**.** Retain the assignments certified in preceding stages. When exactly one formerly idle head has a raw-key coordinate above $1/2$, assign it to its largest raw coordinate. Keep all other idle heads assigned to the sink. Let C^+ be the assigned raw keys, $\mathcal{I}^+$ the remaining idle heads, and $U^+ = [d] \setminus C^+$.

The final stage stops when this assignment has been certified and $E \le d^{-(c+2)}$. For a nonfinal stage, define

$$r_{\min} = \min_{\kappa \in U^+} \max_{h \in \mathcal{I}^+} p_{h\kappa}, \qquad \widetilde{m}_h = \sum_{\kappa \in U^+} p_{h\kappa}, \qquad X^{\mathrm{in}}_{h\kappa} = \frac{p^{\mathrm{in}}_{h\mathrm{BOS}}}{p^{\mathrm{in}}_{h\kappa}}.$$

Stop at the first certified iterate satisfying $E \le E_0$, $\bar{\delta} \le d^8 \delta_*$,

$$\max_{a_h \neq \mathrm{BOS}} (1 - p_{h,a_h}) \le (2 - \tau) r_{\min}, \tag{12}$$

and, for every $h \in \mathcal{I}^+$,

$$(1 - d^{-3}) \frac{\widetilde{m}_h}{|U^+|} \le p_{h\kappa} \le (1 + d^{-3}) \frac{\widetilde{m}_h}{|U^+|} \qquad (\kappa \in U^+), \tag{13}$$

$$p_{hc} \le (2 - \tau)(1 + 2d^{-3}) \frac{\widetilde{m}_h}{|U^+|} \qquad (c \in C^+), \tag{14}$$

$$\frac{p_{h\mathrm{BOS}}}{p_{h\kappa}} \ge (1 - \tau) X^{\mathrm{in}}_{h\kappa} \qquad (\kappa \in U^+). \tag{15}$$

The recorded entry ratios, attention table, and previously observed assignments determine the rule; it does not inspect the secret tuple. If no such iterate occurs by $T_{\max}$, return FAIL.

"Formerly idle" refers to stage entry. Once a new assignment is certified, it is retained during sharpening. Stopping depends only on row-0 attention.

Definition 2 (Conditions at stage entry)**.** For deficit key $t \notin C$, let $U = [d] \setminus C$, $r_h = p_{ht}$, and

$$\zeta = \sum_{h \in \mathcal{I}} \sum_{b \notin \{\mathrm{BOS}, t\}} p_{hb}, \qquad X_{hb} = p_{h\mathrm{BOS}}/p_{hb}.$$

The raw-key uniformity condition with tolerance ς is

$$(1-\varsigma)\frac{\widetilde{m}_h}{|U|} \le p_{hb} \le (1+\varsigma)\frac{\widetilde{m}_h}{|U|} \quad (b \in U), \qquad p_{hc} \le (2-\tau)(1+2\varsigma)\frac{\widetilde{m}_h}{|U|} \quad (c \in C),$$

for every idle head h, where $\widetilde{m}_h = \sum_{b \in U} p_{hb}$. At initialization $C = \varnothing$ and $\varsigma = 0$. At later entries, eqs. (13) and (14) give $\varsigma = d^{-3}$.

B Single-stage convergence analysis

B.1 Concentration and row decoupling

Lemma 1 (All Boolean-character interactions)**.** *With probability at least $1 - p$ over the batch sampled for a stage,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \prod_{j \in S} x_j^i \right| \le \kappa_n \qquad \textit{simultaneously for every nonempty } S \subseteq [d].$$

Proof. For nonempty S, the character values are independent uniform signs. Hoeffding's inequality (Hoeffding, 1963) gives the bound $2e^{-n\kappa_n^2/2}$ for each character. Union bound over fewer than 2^d characters gives failure at most $2^{d+1}e^{-n\kappa_n^2/2} = p$. Repeated indices reduce modulo two; trivial products equal one. □

On this event, a polynomial in Boolean characters with coefficient ℓ_1 norm C has empirical error at most $C\kappa_n$. This bound holds uniformly in its coefficients, and hence along the parameter trajectory.

Lemma 2 (Row decoupling)**.** *Under teacher forcing, row 0 follows GD on*

$$\ell_s(w) = \frac{1}{2n} \left\| \hat{y} - v_{A_s} \right\|_2^2, \qquad \hat{y} = \bigodot_{h=1}^{H} u^{(h)}, \qquad A_s = \{t_1, \ldots, t_{s+1}\}, \tag{16}$$

as if this were the only row loss.

Proof. Ground-truth token values do not depend on the parameters. Every summand of eq. (5) depends on a disjoint row block. The sink scores are fixed, so they introduce no trainable coupling. □

Write $U_{-h} = \bigodot_{g \neq h} u^{(g)}$, $g_{hb} = \partial \ell_s / \partial w_{hb}$, and $a_{hb} = g_{hb}/p_{hb}$. Differentiation gives

$$g_{hb} = \frac{p_{hb}}{n} \langle \hat{y} - v_{A_s}, U_{-h}, v_b - u^{(h)} \rangle, \qquad |a_{hb}| \le 4. \tag{17}$$

Because the sink score is fixed, $X_{hb} = e^{b_h - w_{hb}}$ updates exactly as

$$X_{hb}^{+} = X_{hb} e^{\eta p_{hb} a_{hb}}, \qquad X_{hb}^{+} - X_{hb} = \eta p_{h\mathrm{BOS}} a_{hb} + O(\eta^2 p_{hb}), \qquad |X_{hb}^{+} - X_{hb}| \le 8\eta \tag{18}$$

for $\eta \le 1/16$. In the middle estimate the absolute constant uses $|a_{hb}| \le 4$.

Lemma 3 (Singleton signal)**.** *On the concentration event, suppose previously recruited heads are assigned bijectively to C with error at most ρ, the target is $v_C v_t$ for $t \notin C$, and idle off-sink mass is $\bar{\delta} \le 1/4$. If $k\rho \le 1/4$, then, for every idle head and raw key b,*

$$a_{hb} = -1 - \mathbf{1}\{b = t\} + O(\kappa_n + \bar{\delta} + k\rho).$$

When $\kappa_n + \bar{\delta} + k\rho$ is sufficiently small, the missing key has an order-one normalized-gradient advantage.

Proof. For an idle head, $u^{(h)} = 1 - \theta_h$ with $0 \le \theta_h \le 2m_h$ pointwise. For a previously recruited head, $\|u^{(h)} - v_{\pi(h)}\|_\infty \le 2\rho$. Telescoping products of factors bounded by one therefore gives $(\hat{y} - v_C v_t) U_{-h} = 1 - v_t + O(\bar{\delta} + k\rho)$ pointwise. Also $v_b - u^{(h)} = v_b - 1 + O(m_h)$. The average of $(1 - v_t)(v_b - 1)$ is $-1 - \mathbf{1}\{b = t\} + O(\kappa_n)$ by character orthogonality and Lemma 1. Substitute in eq. (17). □

B.2 Local stability near correct assignments

Lemma 4 (Global smoothness)**.** *For every row-0 input and parameter table, $|\hat{y}| \le 1$, $\|\nabla \hat{y}\|_2 \le 1$, and $\|\nabla^2 \hat{y}\|_{\mathrm{op}} \le 10$. Every empirical or population squared loss with labels in $[-1, 1]$ is globally 21-smooth.*

Proof. For one head, view the token value $V \in \{\pm 1\}$ as random under its attention distribution. Extend a raw-logit direction z by $z_{\mathrm{BOS}} = 0$ and let Z be its random coordinate. Then $Du[z] = \mathrm{Cov}(V, Z)$, so $\|\nabla u\|_2^2 \le 1 - u^2$. A second derivative is the joint third cumulant of V, Z, Z'; centering each bounded coordinate gives $\|\nabla^2 u\|_{\mathrm{op}} \le 8$. Put $q_h = 1 - |u_h|^2$. The squared gradient norm of the product is at most $\sum_h q_h \prod_{g \ne h}(1 - q_g) \le 1$. The diagonal Hessian blocks have operator norm at most 8, and the squared Frobenius norm of the off-diagonal part is at most $\sum_{h \ne g} q_h q_g \prod_{a \notin \{h,g\}}(1 - q_a) \le 2$. These sums are, respectively, the probabilities of exactly one success and twice the probability of exactly two successes for independent Bernoulli variables. Hence the Hessian norm is below $8 + \sqrt{2} < 10$. Finally $\nabla^2 \frac{1}{2}(\hat{y} - y)^2 = \nabla \hat{y} \nabla \hat{y}^\top + (\hat{y} - y)\nabla^2 \hat{y}$ has norm at most 21. □

Lemma 5 (Forward invariance and sharpening)**.** *Fix an assignment $a_h \in \{\mathrm{BOS}, 1, \dots, d\}$ whose raw supports have symmetric difference A_s, and let $E = \sum_h (1 - p_{h,a_h})$. There are universal constants $E_0, c_0, c_1 > 0$ such that, on the concentration event, if $E \le E_0$, $\kappa_n \le c_0$, and $\eta H^2 L \le c_0$, every assigned-to-wrong logit gap is nondecreasing and*

$$\frac{c_1 \eta}{HL} E^3 \le E - E^+ \le c_1^{-1} \eta E^3. \tag{19}$$

In particular the neighborhood is forward invariant, $E^+ \ge E/2$, and $E \le \nu$ after $O(HL/(\eta \nu^2))$ further steps.

Proof. Write $e_h = 1 - p_{h,a_h}$, $v_{\mathrm{BOS}} = 1$, and

$$u^{(h)} = v_{a_h}(1 + \theta_h), \qquad \theta_h = \sum_{b \ne a_h} p_{hb}(v_b v_{a_h} - 1), \quad \|\theta_h\|_\infty \le 2e_h.$$

Factoring out the correct target gives

$$\begin{aligned} v_{A_s}(\hat{y} - v_{A_s}) = \prod_h (1 + \theta_h) - 1 = S + R_2, \qquad \|R_2\|_\infty \le CE^2, \\ S = -E + \sum_{\alpha \ne \varnothing} q_\alpha v_\alpha, \qquad q_\alpha \ge 0, \qquad \sum_\alpha q_\alpha = E. \end{aligned} \tag{20}$$

The coefficients group equal nontrivial characters $v_b v_{a_h}$. Their total mass is E. Thus

$$\frac{1}{n}\|S\|_2^2 = E^2 + \sum_\alpha q_\alpha^2 + O(\kappa_n E^2), \qquad \ell_s \asymp E^2.$$

Let D_{hb} be the derivative for moving probability from the assigned key a_h to a wrong key b. Differentiating the product, including all other heads, gives

$$D_{hb} = E + q_{\mathrm{supp}(b) \triangle \mathrm{supp}(a_h)} + O(\kappa_n E + E^2), \qquad E/2 \le D_{hb} \le 3E, \tag{21}$$

for sufficiently small universal E_0, c_0. Here $\mathrm{supp}(\mathrm{BOS}) = \varnothing$ and $\mathrm{supp}(j) = \{j\}$. Empirical errors scale with E because the residual vanishes at the vertex.

Choose the simplex gauge in which the assigned-key derivative is zero, and put $\bar{D}_h = \sum_{b \ne a_h} p_{hb} D_{hb}$. For trainable coordinates the exact chain rule is

$$g_{h,a_h} = -p_{h,a_h} \bar{D}_h \quad (a_h \ne \mathrm{BOS}), \qquad g_{hb} = p_{hb}(D_{hb} - \bar{D}_h) \quad (b \ne a_h, \mathrm{BOS}). \tag{22}$$

For a raw assigned key and a wrong raw key, the assigned gap increases by $\eta[(p_{h,a_h} - p_{hb})\bar{D}_h + p_{hb} D_{hb}] \ge 0$. The assigned-to-sink gap increases by $\eta p_{h,a_h} \bar{D}_h \ge 0$. For a sink assignment, $\bar{D}_h \le 3Ee_h$ and $g_{hb} \ge p_{hb} E/4 > 0$, so its sink-to-raw gaps also increase. This proves forward invariance.

For a raw-assigned head, the first-order decrease of e_h is

$$\eta p_{h,a_h}\left[p_{h,a_h}e_h\bar{D}_h+\sum_{b\notin\{a_h,\mathrm{BOS}\}}p_{hb}^2(D_{hb}-\bar{D}_h)\right].$$

Both terms are nonnegative: $D_{hb}-\bar{D}_h \geq E(1/2-3e_h) \geq E/4$ for sufficiently small E_0. Since $\bar{D}_h \geq Ee_h/2$, the first term is at least $c\eta Ee_h^2$. For a sink-assigned head the decrease is $\eta p_{h\mathrm{BOS}}\sum_{b\neq\mathrm{BOS}}p_{hb}^2(D_{hb}-\bar{D}_h) \geq c\eta E\sum_{b\neq\mathrm{BOS}}p_{hb}^2$. Summation and Cauchy–Schwarz over at most HL tail probabilities yield a total decrease at least $c\eta E^3/(HL)$, and the corresponding upper bound is $C\eta E\sum_h e_h^2 \leq C\eta E^3$.

To bound the discrete remainder, note that $\|g_h\|_\infty \leq CEe_h$. The bound $|D^2e_h[z,z]| \leq C\|z\|_\infty^2$ follows from the identity

$$D^2p_{h,a_h}[z,z]=p_{h,a_h}[(z_{a_h}-\bar{z})^2-\mathrm{Var}_p(z)],\qquad z_{\mathrm{BOS}}=0.$$

The summed discrete remainder is at most $C\eta^2E^2\sum_h e_h^2 \leq C\eta^2E^4$ and is absorbed by the lower bound when ηHL is sufficiently small. Equation (19) follows. It gives $(E^+)^{-2} \geq E^{-2}+c\eta/(HL)$ and, with a smaller step size if needed, $E^+ \geq E/2$. □

A raw-assigned head with sink probability p_s, wrong-raw tail t_h, and $t_h \leq \chi p_s$ has $Y_h = p_{h,a_h}/p_s$ satisfying $Y_he_h \leq 1+\chi$. In the same vertex neighborhood, every wrong raw logit decreases, since $D_{hb}-\bar{D}_h \geq E/4$. Hence $t_h/p_s=\sum_{b\notin\{a_h,\mathrm{BOS}\}}e^{w_{hb}-b_h}$ is nonincreasing there.

B.3 Reduced dynamics and head competition

Fix a one-deficit stage with target v_Cv_t. Let $r_h = p_{ht}$ for idle heads, $\zeta_h=\sum_{b\notin\{\mathrm{BOS},t\}}p_{hb}$, and $\zeta=\sum_h\zeta_h$. Write $F=\prod_{h\in\mathcal{I}}(1-2r_h)$ and $F_{-h}=\prod_{g\in\mathcal{I}\setminus\{h\}}(1-2r_g)$. The empty product is one.

Lemma 6 (Reduction with the target mass retained)**.** *On the concentration event, suppose the errors of previously recruited heads are at most ρ and $k\rho \leq 1/4$. Uniformly for idle heads,*

$$\ell_s=\tfrac{1}{4}(F+1)^2+O(\kappa_n+\zeta+k\rho), \tag{23}$$

$$-a_{ht}=(F+1)F_{-h}(1-r_h)+O(\kappa_n+\zeta+k\rho), \tag{24}$$

$$-a_{hb}=(F+1)F_{-h}(\tfrac{1}{2}-r_h)+O(\kappa_n+\zeta+k\rho)\qquad(b\notin\{\mathrm{BOS},t\}). \tag{25}$$

The exact target flow is

$$\dot{r}_h=-r_h^2(1-r_h)a_{ht}+r_h\sum_{b\notin\{\mathrm{BOS},t\}}p_{hb}^2a_{hb}. \tag{26}$$

If, in addition, $p_{hb} \leq C_{\mathrm{diff}}r_h$ for every idle distractor, then before the first crossing $r_h \leq 1/2$,

$$r_h^+=r_h+\eta(F+1)F_{-h}r_h^2(1-r_h)^2+O\big(\eta(\kappa_n+\zeta+k\rho+\eta)r_h^2\big). \tag{27}$$

The constant in eq. (27) *may depend on the absolute coordinate bound C_{diff}.*

Proof. Replace each idle output by $\bar{u}_h=1-r_h(1-v_t)$; the pointwise error is at most $2\zeta_h$. Replace the product of previously recruited heads' outputs by v_C; the error is at most $2k\rho$ by telescoping. The resulting residual and all leave-one-out products differ by $O(\zeta+k\rho)$ in max-norm. For the two-key products $\bar{Q},\bar{Q}_{-h}$,

$$\bar{Q}-v_t=(F+1)(1-v_t)/2,\qquad(\bar{Q}-v_t)\bar{Q}_{-h}=(F+1)F_{-h}(1-v_t)/2.$$

Averaging their square gives eq. (23). Pairing the second identity with $v_t-\bar{u}_h=-(1-r_h)(1-v_t)$ gives eq. (24). For $b\neq t,\mathrm{BOS}$, its pairing with $v_b-\bar{u}_h$ has expectation $-(F+1)F_{-h}(1/2-r_h)$: terms involving v_b, v_t, and v_bv_t are nontrivial characters. The empirical replacement costs $O(\kappa_n)$ because the coefficient norms are absolute. This gives eq. (25).

Differentiating r_h with the sink score fixed gives eq. (26). Since $\sum_b p_{hb}^2 \leq C_{\mathrm{diff}}r_h\zeta_h$ and $|a_{hb}| \leq 4$, the second term is $O(r_h^2\zeta)$. For the discrete remainder, $\|g_h\|_\infty \leq Cr_h$ on the coordinate bound, and $|D^2r_h[z,z]| \leq Cr_h\|z\|_\infty^2$ along a sufficiently short step. Thus one softmax update has remainder $O(\eta^2r_h^3)$, smaller than the stated error. □

Lemma 7 (Distractors before the crossing)**.** *Consider a stopped pre-crossing interval of at most $C_0 G(r_1^0)/\eta$ steps, where head 1 is the initial leader and $G(r) = 1/[r(1-r)]$. Suppose entry has tolerance $\varsigma \le c\tau$, off-sink mass at most $1/4$, and $r_h \le 1/2$ on the interval. Assume old errors stay below ρ, follower mass stays below $1/8$, and*

$$\zeta \le \bar{\zeta} = 2C_\zeta(\tau^{-1}\zeta^0 + \kappa_n + k\rho + \eta).$$

Put $\epsilon = C(\bar{\zeta} + \kappa_n + k\rho + \eta)$. If

$$\epsilon G(r_1^0) \le c\tau \min_h X_{ht}^0, \tag{28}$$

then, for sufficiently small c (depending on the fixed horizon constant), throughout the interval

$$p_{hb} \le 2r_h \quad (b \ne t, \mathrm{BOS}), \qquad \zeta \le \bar{\zeta}/2.$$

An unclaimed distractor satisfies $X_{hb} \ge 0.4X_{hb}^0$; every claimed distractor satisfies $X_{hb} \ge c\tau X_{hb}^0$. For two unclaimed distractors of one head, their reciprocal increments differ cumulatively by at most $C\epsilon G(r_1^0)$.

Proof. Before the crossing, $(F+1)F_{-h} \ge 0$, so eqs. (24) and (25) give

$$-a_{hb} \le \tfrac{1}{2}(-a_{ht}) + C\epsilon.$$

Multiplying by the common sink mass in eq. (18) and summing gives

$$X_{hb}^m \ge X_{hb}^0 - \tfrac{1}{2}(X_{ht}^0 - X_{ht}^m) - C\epsilon G(r_1^0). \tag{29}$$

For an unclaimed distractor, $X_{hb}^0 \ge (1 - O(\varsigma))X_{ht}^0$. For a claimed one, the entry cap gives $X_{hb}^0 \ge (1/2 + c'\tau)X_{ht}^0$ after choosing the tolerance constant small enough. Equation (28) then leaves $X_{hb}^m \ge X_{ht}^m/2$ and the stronger absolute bounds in the statement. Consequently, $p_{hb} \le 2r_h$, and the distractor odds increase by at most $C\tau^{-1}$. Entry sink masses are at least $3/4$, so summing the odds proves $\zeta \le C\tau^{-1}\zeta^0 \le \bar{\zeta}/2$ after the fixed constants are chosen. For two unclaimed distractors, the leading term in eq. (25) is identical. Subtract the reciprocal updates: the per-step difference is $O(\eta\epsilon)$, proving the last assertion. These are strict improvements of the stopped barriers. □

Lemma 8 (A discrete half-rate comparison)**.** *Let $C_1 \ge 1$. Suppose $0 < e_m \le 1/2$ and $0 < r_m \le 1/2$ before the final step, which may cross $1/2$. Suppose the interval has at most $C_1/(\eta r_0)$ steps, and, with $b_m = (1-r_m)(1-2r_m)$ and $\upsilon \ge 0$,*

$$e_{m+1} \le e_m + \eta b_m e_m^2 + C_1\eta\upsilon e_m^2, \qquad |e_{m+1} - e_m| \le C_1\eta e_m^2,$$
$$r_{m+1} \ge r_m + 2\eta r_m^2(1-r_m)^3 - C_1\eta\upsilon r_m^2, \qquad |r_{m+1} - r_m| \le C_1\eta r_m^2.$$

For sufficiently small $C_1\eta$, every prefix obeys

$$\frac{1}{e_m} \ge \frac{1}{e_0} - \tfrac{1}{2}\{G(r_0) - 4\} - \frac{CC_1^3(\upsilon + \eta)}{r_0}. \tag{30}$$

Proof. The reciprocal identity gives $e_{m+1}^{-1} \ge e_m^{-1} - (e_{m+1} - e_m)/e_m^2 \ge e_m^{-1} - \eta b_m - C_1\eta\upsilon$. For $\varphi = G/2$, use $-\varphi'(r) = (1-2r)/(2r^2(1-r)^2)$ and $|\varphi''(r)| \le C/r^3$ on the one-step interval. The target bounds imply $\varphi(r_m) - \varphi(r_{m+1}) \ge \eta b_m - CC_1^2\eta(\upsilon + \eta)$. Sum, use $G \ge 4$ also at the possible crossing endpoint, and then use the interval length. □

Lemma 9 (One head is recruited and older heads survive)**.** *There are universal constants $C_L, C_0, c_0 > 0$ with the following property. At stage entry, assume the conditions of Definition 2 with $\varsigma \le c_0\tau$ and $\bar{\delta}^0 \le 1/4$. Let r_1^0 be the unique largest target mass among idle heads, put*

$$c_i = G(r_i^0) - G(r_1^0) > 0, \qquad \Sigma = \sum_{i \ne 1}(c_i + 4)^{-1}, \qquad \rho_{\max} = 2C_L\tau^{-1}r_1^0,$$

and set $\epsilon = C(\kappa_n + k\rho_{\max} + \tau^{-1}\zeta^0 + \eta)$ with a sufficiently large absolute C. Suppose each old head has $e_h^0 \le (2-\tau)r_1^0$ and $t_h^0 \le \chi p_{h\mathrm{BOS}}^0/2$, and

$$r_1^0 \le \tau/8, \qquad H\tau^{-1}r_1^0 \le c_0, \qquad d\chi/\tau \le c_0, \tag{31}$$

$$C\big(kr_1^0/\tau + \Sigma + \tau^{-1}\zeta^0 + \kappa_n + \eta + \chi\big) \le c_0\tau, \tag{32}$$

$$\epsilon G(r_1^0) \le c_0 \min\{\tau G(r_1^0), \min_{i \ne 1} c_i\}. \tag{33}$$

The inner minimum is $+\infty$ with one idle head. Then exactly the initial leader crosses $1/2$. All other idle heads remain below $1/2$, and the enlarged assignment enters the E_0 neighborhood in at most $C_0 G(r_1^0)/\eta + CH/\eta$ steps. Old errors stay below $C_L \tau^{-1} r_1^0$ and their wrong-raw/sink ratios grow by at most

$$\exp\{C\chi/\tau + CH\chi\tau^{-1} r_1^0\}. \tag{34}$$

The new head has wrong-raw/sink ratio at most $C\tau^{-1}\bar{\delta}^0$ at the crossing and at most that bound times $e^{C\epsilon}$ on entry to the vertex neighborhood. After entry, all assigned gaps increase and these tail ratios decrease.

Proof. We control head competition and old assignments on the same stopped interval. Before the first crossing, stop when any old error reaches $\rho_{\max}$, an old tail ratio reaches χ, the follower sum reaches $1/8$, a distractor barrier reaches $\bar{\zeta}$ as in Lemma 7, or the pre-crossing horizon $C_{\mathrm{pre}} G(r_1^0)/\eta$ is reached. Constants are chosen in the order: fixed large C_{pre} and transition constant; large C_L; then sufficiently small c_0; finally a horizon constant covering both parts. All estimates below hold for every prefix of this stopped interval.

Pre-crossing competition. Entry sink masses exceed $3/4$, so $\min_h X_{ht}^0 \geq 3/(4r_1^0) \geq (3/8)G(r_1^0)$. The first term of eq. (33) verifies eq. (28). Thus Lemma 7 applies, and eq. (27) has error at most $C\eta\epsilon r_h^2(1-r_h)^2$. Since $G'(r) = -(1-2r)/(r^2(1-r)^2)$,

$$G(r_h^{m+1}) - G(r_h^m) = -\eta(F^m + 1)F^m + \theta_h^m, \qquad |\theta_h^m| \leq C\eta\epsilon + C\eta^2 r_h^m. \tag{35}$$

Subtract two heads and sum. Equation (33) preserves at least half of every leader–follower gap. Therefore

$$\sum_{i \neq 1} r_i^m \leq \sum_{i \neq 1} \frac{2}{4 + c_i/2} \leq 4\Sigma \leq 1/16, \tag{36}$$

a strict improvement. While $r_1 \leq 1/4$, its reciprocal decreases at an absolute rate; from $1/4$ to $1/2$, its target probability increases at an absolute multiple of η because $F_{-1} \geq 7/8$. It crosses within the pre-crossing horizon. A discrete jump cannot make a follower cross: raw probabilities change by a factor at most $e^{8\eta}$ per step. If the leader lands exactly at $1/2$, its next increment is at least $\eta/32$ after absorbing the error. Distractor masses are $O(\epsilon)$, so the newly certified coordinate is the deficit key t.

Old-head error up to the crossing. For an old head assigned to a, let $e = 1 - p_a$, $p_s = p_{\mathrm{BOS}}$, and $t_h = \sum_{b \notin \{a, \mathrm{BOS}\}} p_b$. On the stopped localization bound, $t_h \leq \chi p_s \leq \chi e$. Collapse this tail to the sink and keep only the leader among idle heads. After factoring out the other assigned characters, the ideal loss is

$$L(e, r) = \tfrac{1}{2}\{e^2(1-r)^2 + (1-e)^2(1-r)^2 + e^2 r^2 + [(1-e)r - 1]^2\}. \tag{37}$$

Its derivatives are

$$L_e = -(1-r)(1-2r) + 2e[(1-r)^2 + r^2], \qquad L_r = -(1-e) + (2r-1)[e^2 + (1-e)^2].$$

In the ideal system $\dot{e} = -e^2(1-e)^2 L_e$ and $\dot{r} = -r^2(1-r)^2 L_r$. Consequently, $\dot{e} \leq e^2(1-r)(1-2r)$ and $\dot{r} \geq 2r^2(1-r)^3 - Cer^2$ for $r \leq 1/2$.

Let h denote the old head and h_1 the leading idle head. The normalized directions are

$$\widetilde{D}_e = -\frac{a_{ha}}{e}, \qquad \widetilde{D}_r = \frac{a_{h_1 t}}{1-r}.$$

In the ideal two-head system they equal L_e and L_r, respectively. In the actual system, the old head differs from $(1-e)v_a + e$ by at most $2t_h$. Its assigned direction $(v_a - u_h)/e$ differs from $v_a - 1$ by at most $4t_h/e \leq 4\chi$. The leading head's direction, divided by $1 - r \geq 1/2$, incurs $O(\zeta)$ error. Telescoping the other head outputs contributes $O(k\rho_{\max} + \Sigma + \zeta)$. The ideal inner products contain only $1, v_a, v_t, v_a v_t$ and have bounded coefficient norms, so their empirical error is $O(\kappa_n)$. Therefore

$$|\widetilde{D}_e - L_e| + |\widetilde{D}_r - L_r| \leq C(\kappa_n + k\rho_{\max} + \Sigma + \zeta + \chi). \tag{38}$$

The exact old-error flow is

$$\dot{e} = (1-e)^2 e\, a_a - (1-e) \sum_{b \notin \{a, \mathrm{BOS}\}} p_b^2 a_b. \tag{39}$$

Here $a_a = -eL_e + O(e\upsilon)$ and $\sum_b p_b^2 \le \chi^2 e^2$, with $\upsilon = C(\kappa_n + k\rho_{\max} + \Sigma + \tau^{-1}\zeta^0 + \chi + \eta)$. Taylor expansion therefore gives the one-sided recursions in Lemma 8. Their two-sided bounds follow from $|v_a - u_h| \le 2e$, which bounds every old assigned gap change by $8\eta e$, and from the coordinate bound $p_{hb} \le 2r_h$ for the leader. The leader grows by at least $c\eta r^2$ before $1/2$, so the interval has at most $C/(\eta r_1^0)$ steps. Indeed, $r^+ \ge r + c\eta r^2$ implies $1/r - 1/r^+ \ge c\eta/(1 + c\eta r) \ge c\eta/2$. The half-rate lemma yields

$$\frac{1}{e_m} \ge \frac{1}{e_0} - \tfrac{1}{2}\{G(r_1^0) - 4\} - \frac{C(\upsilon + \eta)}{r_1^0}.$$

Since $e_0 \le (2 - \tau)r_1^0$ and $r_1^0 \le \tau/8$,

$$\frac{1}{e_0} - \tfrac{1}{2}G(r_1^0) + 2 \ge \frac{\tau}{8r_1^0} + 2.$$

Equation (32) absorbs the perturbation. Increasing C_L if necessary gives $e_m \le (C_L/2)\tau^{-1}r_1^0$, strictly inside the stopped old-error barrier.

Old-tail localization. For a wrong raw coordinate of an old head,

$$|\Delta \log(p_b/p_s)| = \eta|g_b| \le 4\eta p_b \le 4\eta\chi e.$$

The preceding error bound and the pre-crossing duration give logarithmic growth at most $C\chi/\tau$. Starting from $t_h/p_s \le \chi/2$, the ratio stays below $3\chi/4$ after choosing the constants in eq. (31) small enough. This also excludes its stopped boundary. Hence the first stopped exit is the unique target crossing.

Post-crossing transition. Restart the stopped comparison at the crossing, with barriers $p_{hb} \le 3r_h$, $\zeta \le 2\bar{\zeta}$, follower sum at most $1/8$, the same old-error and localization barriers, and a leader relapse barrier $r_1 \ge 5/8$ after it first reaches $3/4$. Stop at vertex entry or after $C_{\mathrm{tr}}H/\eta$ steps. At the crossing, eq. (29) gives

$$X_{hb} \ge \max\{X_{ht}/2, c\tau/r_1^0\} \quad (b \ne t, \mathrm{BOS}).$$

All reciprocals change by $O(H)$ during this stopped transition. Since $H\tau^{-1}r_1^0 \le c_0$, this preserves $X_{hb} \ge X_{ht}/3$. At the crossing the leader sink mass is at least $1/3$ and follower sink masses exceed $7/8 - O(\epsilon)$. Converting distractor mass to odds costs at most a factor 3; the reciprocal lower bound makes the additional relative change $1 + O(c_0)$. Consequently $\zeta < 2\bar{\zeta}$ throughout the stopped transition.

To control signs, put $q = 1 - r_1$, $S = \sum_{i\ne 1} r_i$, and $P = \prod_{i\ne 1}(1 - 2r_i) \ge 3/4$. Then

$$F + 1 = 1 - (1 - 2q)P \ge c(q + S).$$

Before vertex entry, old errors and distractors total at most $E_0/8$ by the smallness assumptions, so $q + S \ge E_0/2$. Equation (25) implies, for every leader or follower distractor,

$$a_{hb} \ge cE_0(2r_1 - 1) - C\epsilon.$$

Thus every wrong-raw/sink log-odds is nonincreasing outside the layer $|2r_1 - 1| \le C_{\mathrm{bl}}\epsilon$. In that layer the leader advances at least $\eta/32$ per step; the layer lasts $O((\epsilon + \eta)/\eta)$ steps, and follower growth is bounded by $C\eta\epsilon r_i^2$. Thereafter the leader reaches $3/4$ in $O(1/\eta)$ steps while $S \le C\Sigma$ and every follower remains below $1/2$.

The leader's wrong-raw/sink odds are $O(\epsilon)$ at the crossing, change by a factor $e^{O(\epsilon)}$ through the layer, and decrease afterward. Hence its distractor mass is $O(\epsilon q)$. Once $r_1 \ge 3/4$, the retained-target identities and the favorable nonnegative distractor term in eq. (26) give

$$\begin{aligned} r_i^+ &\le r_i - c\eta(q + S)r_i^2 + C\eta\epsilon r_i^2 \quad (i \ne 1), \\ q^+ &\le q - c\eta q^2(q + S) + C\eta\epsilon q. \end{aligned} \tag{40}$$

For the leader, its target normalized gradient is $-(F + 1)F_{-1}q + O(\epsilon)$; the discrete remainder is $O(\eta^2 q)$ and is included because $\eta \le \epsilon$. When $S \ge E_0/8$, use $\sum_i r_i^2 \ge S^2/H$ to reduce S below that threshold in $O(H/\eta)$ steps. When $q \ge E_0/8$, its negative term reduces q below that threshold in $O(1/\eta)$ steps. Dropping the negative terms gives $q^+ \le (1 + C\eta\epsilon)q$ and $S^+ \le S + C\eta\epsilon S^2$. Since $H\epsilon = o(1)$ under eq. (32), either quantity stays below $E_0/7$ after its first threshold crossing, and the

leader cannot relapse to $5/8$. Both are therefore below $E_0/7$ by the end of the transition. Adding the old errors and distractors enters the E_0 neighborhood. All stopped barriers have been improved.

An old error obeys $e^+ \le e + C\eta e^2$ without the half-rate refinement. Over the $O(H/\eta)$ transition it changes by at most a factor $1 + CH\tau^{-1}r_1^0 < 9/8$, leaving $e < C_L\tau^{-1}r_1^0$. Old-tail log growth adds at most $CH\chi\tau^{-1}r_1^0$, proving eq. (34). For the new head, Lemma 7 bounds its off-target odds by $C\tau^{-1}\bar{\delta}^0$ at the crossing. Only the small crossing layer can increase them before vertex entry, by the factor stated. Finally Lemma 5 supplies all post-entry monotonicity conclusions. □

C PERSISTENCE ACROSS DELETION STAGES

We now show that each completed stage supplies the hypotheses of Lemma 9 for the next stage, and that the total training cost remains polynomial.

Head relabeling. Permuting heads, together with their fixed sink scores, leaves the model, loss, and update unchanged. We may therefore relabel them so that

$$\Lambda_1 < \cdots < \Lambda_H, \qquad \Lambda_{h+1} - \Lambda_h \ge g_* \ge d^{B-K}. \tag{41}$$

Initially, $X_{h\kappa} = \Lambda_h$ for every raw key and

$$G(p^0_{h\kappa}) = \Lambda_h + d + 1 + \frac{1}{\Lambda_h + d - 1}. \tag{42}$$

Thus an adjacent G gap is at least half the corresponding Λ gap for large d. In particular, for any remaining terminal segment of the ordered heads,

$$\sum_{i \ne 1} (G(r_i^0) - G(r_1^0) + 4)^{-1} \le C\frac{1 + \log d}{g_*}. \tag{43}$$

The same estimate will follow at later entries from preservation of pairwise reciprocal gaps.

Lemma 10 (Uniform bounds across stages)**.** *Under eqs.* (6)*,* (7) *and* (10) *and Specification 1, on the concentration event for every stage, all nonfinal stages reach Definition 1. At every stage entry, head $h_s = s + 1$ is the unique largest-attention idle head on every unclaimed key, and*

$$\frac{1}{4\mathfrak{B}} \le r_h^0 \le \frac{3}{\Lambda_1}, \qquad G(r^0_{h_s}) \le 8\mathfrak{B}, \qquad G(r_i^0) - G(r^0_{h_s}) \ge \tfrac{1}{2}(\Lambda_i - \Lambda_{h_s}), \qquad \mathfrak{B} = d^6\Lambda_H.$$

The last inequality concerns the remaining followers $i > h_s$. The entry tolerance is at most d^{-3}, $\bar{\delta} \le d^8\delta_*$, and every old head has

$$e_h \le (2 - \tau)r_{\min}, \qquad t_h \le (\chi/2)p_{h\mathrm{BOS}}.$$

All hypotheses of Lemma 9 hold. Each capture and nonfinal recovery, and the final capture before accuracy sharpening, take fewer than $d^R/(2\eta)$ steps for sufficiently large d.

Proof. Put

$$\mathfrak{B} = d^6\Lambda_H, \qquad N = d^{2K+30}, \qquad X_* := \Lambda_1/32, \qquad \theta := \kappa_n + \eta + \chi + d^{10}\delta_* + d/g_*. \tag{44}$$

For an old head assigned to a_ℓ, use $Y_\ell = p_{\ell a_\ell}/p_{\ell\mathrm{BOS}}$ and $J_\ell = 1/e_\ell$. Localization gives

$$Y_\ell = (J_\ell - 1)(1 + \lambda_\ell), \qquad 0 \le \lambda_\ell = t_\ell/p_{\ell\mathrm{BOS}} \le \chi. \tag{45}$$

A common reciprocal coordinate. At stage s, let $h_s = s + 1$ be the proposed leader, $r_s = r^{\mathrm{in}}_{h_s}$, and $Z_s = X^{\mathrm{in}}_{h_s,t}$ for the current deficit t. During the pre-crossing part define

$$v_m = \tfrac{\eta}{2}(F^m + 1)F^m \ge 0, \qquad D_s(m) = \sum_{a<m} v_a, \tag{46}$$

where the sum is truncated at the first crossing. Write D_s for its value there. On every stopped prefix where the capture bounds hold, eq. (35) and eqs. (18) and (25) give

$$\begin{aligned} G(r_h^{m+1}) - G(r_h^m) &= -2v_m + R_h^{G,m}, \\ X_{h\kappa}^{m+1} - X_{h\kappa}^m &= -v_m + R_{h\kappa}^{X,\mathrm{race},m} \quad (\kappa \ne t, \mathrm{BOS}), \\ [-v_m - (J_\ell^{m+1} - J_\ell^m)]_+ &= R_\ell^{J,m}. \end{aligned} \tag{47}$$

The second residual includes the leading correction $r_h^m v_m$: indeed its leading increment is $-(1 - r_h^m)v_m$. This correction is summable. Since the leader grows at least $c\eta r_{h_s}^2$ before the crossing and $r_h \le r_{h_s}$,

$$\sum_{\text{race}} \eta r_h^m \le C \log \mathfrak{B}. \tag{48}$$

Indeed, $r_{h_s}^+ \ge r_{h_s} + c\eta r_{h_s}^2$ implies $\log r_{h_s}^+ - \log r_{h_s} \ge c\eta r_{h_s}/2$ for small η. Summing up to the crossing gives the logarithmic bound. For the third residual, the old-error proof gives adverse reciprocal drift at most $\eta b_m + C\eta\theta$, with $b_m = (1 - r_{h_s})(1 - 2r_{h_s})$. Writing $F = (1 - 2r_{h_s})P$ and $|1 - P| \le C\Sigma$ shows $|\eta b_m - v_m| \le C\eta\Sigma$. Consequently

$$\max_h \sum |R_h^{G,m}| + \max_{h,\kappa} \sum |R_{h\kappa}^{X,\text{race},m}| + \max_\ell \sum R_\ell^{J,m} \le C\theta Z_s + C \log \mathfrak{B}. \tag{49}$$

The maxima are over coordinates; the sums are over training steps.

During sharpening, the vertex expansion gives the exact residual definitions

$$\begin{aligned} X_{h\kappa}^{m+1} - X_{h\kappa}^m &= \eta(E^m + q_{\{\kappa\}}^m) + R_{h\kappa}^{X,m}, \\ Y_\ell^{m+1} - Y_\ell^m &= \eta(E^m + q_{\{a_\ell\}}^m) + R_\ell^{Y,m}. \end{aligned} \tag{50}$$

For an unclaimed raw key, $q_{\{\kappa\}} = S_\kappa := \sum_{h \in \mathcal{I}} p_{h\kappa}$: no raw-assigned head can create that singleton character. For a claimed key c, $q_{\{c\}} \ge p_{hc}$ for an idle head and $q_{\{a_\ell\}} \ge p_{\ell\text{BOS}} \ge (1 - \chi)e_\ell$. The residuals obey

$$\begin{aligned} |R_{h\kappa}^{X,m}| &\le C\eta[(m_h + \kappa_n + \eta)E^m + (E^m)^2], \\ |R_\ell^{Y,m}| &\le C\eta[(\chi + \kappa_n + \eta)E^m + (E^m)^2]. \end{aligned} \tag{51}$$

To verify the first line, use $X_{h\kappa} g_{h\kappa} = p_{h\text{BOS}}(D_{h\kappa} - \bar{D}_h)$ in eq. (18), with $p_{h\text{BOS}} = 1 - m_h$ and $\bar{D}_h = O(m_h E)$. For the second line, $Y_\ell^+ = Y_\ell e^{-\eta g_{\ell a_\ell}}$ and $-Y_\ell g_{\ell a_\ell} = Y_\ell p_{\ell a_\ell} \bar{D}_\ell$. The sink term is $p_{\ell a_\ell}^2 (E + q_{\{a_\ell\}})$; localization bounds the other terms by $O(\chi E)$. The quadratic vertex and exponential-update errors give the remaining displayed terms.

Define $\mathcal{E}_s$ to be a fixed sufficiently large constant times: the three maxima of cumulative residuals in eq. (49); the maximum prefix changes of each X and Y during the post-crossing transition; the two maxima of cumulative absolute residuals in eq. (51); and the allowances $\chi\mathfrak{B} + d + \log \mathfrak{B} + \eta$. Every constituent is truncated at the current iterate. Put $\mathcal{E}_{\le s} = \sum_{j<s} \mathcal{E}_j + \mathcal{E}_s$. This cumulative error is nondecreasing. The $\chi\mathfrak{B}$ allowance covers the adverse J-to-Y conversion in eq. (45).

For surviving idle heads on a future key, the leading increments in eqs. (47) and (50) are common across heads. Subtracting the updates and accumulating residuals and transition changes over completed stages and the current prefix gives

$$|(X_{h\kappa} - X_{h'\kappa}) - (\Lambda_h - \Lambda_{h'})| \le C\mathcal{E}_{\le s}.$$

Induction hypotheses. During capture, a future key means a raw key other than the already claimed keys and the current deficit; after capture it means any still-unclaimed raw key. Conditions on future coordinates are vacuous when no such key remains. At stage entry, the full unclaimed-key uniformity is inherited from the preceding stopping rule. This induction ends at the final capture; the final accuracy iterations are treated separately in Section D. Sharpening intervals below refer only to nonfinal stages.

Consider the maximal actual prefix on which all current Z_s, all idle raw $X_{h\kappa}$, and assigned Y_ℓ are at most $2\mathfrak{B}$, future $X_{h\kappa} \ge X_*$, $\mathcal{E}_{\le s} \le 2N$, adjacent future head gaps exceed half their initial gaps, and

$$|X_{h\kappa} - X_{h\kappa'}| \le \tfrac{1}{2} d^{-3} X_* \quad \text{for two future keys of one idle head.} \tag{52}$$

All inequalities are strict at initialization. For a nonfinal stage, extend the fixed-target trajectory only for comparison until the buffered recovery conditions and $E \le c\tau^2$ hold. The estimates apply to prefixes through the possible first-exit step; at deletion, we use the actual stopping iterate. The bounds $|X^+ - X| \le 8\eta$, $|Y^+ - Y| \le C\eta$, and $E^+ \ge E/2$ in the vertex neighborhood absorb endpoint overshoots.

Inductively, suppose completed stages satisfy the stopping rule and cumulative localization. For every head and raw key still idle and unclaimed at entry to stage s, the actual relative-floor clauses telescope:

$$X_{h\kappa}^{\mathrm{in}} \geq (1-\tau)^s \Lambda_h \geq \tfrac{3}{4}\Lambda_h. \tag{53}$$

This gives an absolute scale bound even when adjacent initial gaps are much smaller than Λ_1.

At an entry, raw-key uniformity and the claimed-coordinate cap imply $G(r_h) = X_{ht} + O(d)$: exactly, $G(r_h) - X_{ht} = 1 + 1/(1-r_h) + \zeta_h/r_h$, and every raw distractor is at most a fixed multiple of r_h. Preserved gaps therefore give

$$G(r_i) - G(r_{h_s}) \geq \tfrac{1}{4}(\Lambda_i - \Lambda_{h_s}), \qquad \Sigma \leq C(1+\log d)/g_*. \tag{54}$$

We check probability dominance below. Once it holds, $r_{\min} \leq r_{h_s,t}$, so the prior balance condition is the old-error premise of Lemma 9. With $r_{h_s} \leq C/\Lambda_1$ and $\bar{\delta}^0 \leq d^8\delta_*$,

$$\epsilon_s \leq C(\kappa_n + \eta + d^{10}\delta_*). \tag{55}$$

The choices of exponents ensure $\epsilon_s \mathfrak{B} = o(\tau X_*)$ and $o(g_*)$, $\Sigma = o(\tau)$, and all the other smallness conditions in eqs. (31) and (32). These are checked quantitatively below. Thus Lemma 9 establishes the next capture on this stopped prefix.

The new head has wrong-raw/sink ratio at most $C\tau^{-1}d^8\delta_*\, e^{C\epsilon_s} \leq \chi/4$ at vertex entry for sufficiently large d. Across at most d later captures, eq. (34) bounds cumulative logarithmic growth by

$$Cd\chi/\tau + CdH\chi\tau^{-1}/\Lambda_1 = o(1).$$

This is less than $\log(4/3)$ for large d. Vertex intervals only decrease these ratios. It follows prefixwise that every previously recruited head has $t_h/p_{h\mathrm{BOS}} \leq \chi/2$. This supplies localization at the next entry.

Ordering of idle heads. An idle follower throughout a capture and transition has

$$1 - p_{h\mathrm{BOS}} \leq C(\Sigma + \tau^{-1}\zeta^0 + k\rho_{\max} + \kappa_n + \eta) \leq C(d/g_* + d^{10}\delta_* + \kappa_n + \eta). \tag{56}$$

The post-crossing follower recursion preserves $S \leq C\Sigma$ because its possible growth is at most $C\eta\epsilon_s S^2$ over $O(H/\eta)$ steps. In the vertex neighborhood, all raw-to-sink odds of an idle head decrease. Its total off-sink mass therefore decreases, so eq. (56) persists. At stage entry, this bound is inherited from initialization or the preceding completed stage. During capture, eq. (56) applies to the remaining followers. On the stopped reciprocal bands, any slower idle head and the least-scale idle head h^+ satisfy

$$\frac{p_{h^+\kappa}}{p_{h\kappa}} = \frac{p_{h^+\mathrm{BOS}}}{p_{h\mathrm{BOS}}}\frac{X_{h\kappa}}{X_{h^+\kappa}} \geq (1 - o(g_*/\mathfrak{B}))(1 + g_*/(4\mathfrak{B})) > 1.$$

Here eq. (56) is $o(g_*/\mathfrak{B})$ under the exponent choices. Thus the same least-scale idle head realizes the maximum on every future key. This verifies the leader used above and the $r_{\min}$ conversion for the next stage.

Change during capture. Sum eq. (47) to the first crossing. There $G(r_{h_s}) = 4 + O(\eta^2)$, and at entry $G(r_{h_s}) = Z_s + O(d)$. Therefore

$$2D_s = Z_s + O(d + \mathcal{E}_s), \qquad X_{h\kappa}^{\mathrm{cross}} = X_{h\kappa}^{\mathrm{in}} - D_s + O(\mathcal{E}_s) \quad (\kappa \neq t, \mathrm{BOS}). \tag{57}$$

For a target coordinate of a follower, the decrement is instead $2D_s + O(d + \mathcal{E}_s)$, by the first line of eq. (47) and the same G–odds conversion. Old heads obey

$$Y_\ell^{\mathrm{cross}} \geq Y_\ell^{\mathrm{in}} - D_s - O(\mathcal{E}_s). \tag{58}$$

Indeed J has at most the adverse loss recorded in eq. (47), and $Y^{\mathrm{in}} \leq J^{\mathrm{in}} - 1 + \chi J^{\mathrm{in}}$ whereas $Y^{\mathrm{cross}} \geq J^{\mathrm{cross}} - 1$. The conversion cost is already counted. Transition changes of X, Y are also counted.

Recovery before the next deletion. For a nonfinal stage, set $u = d - s - 1 \geq 1$, $h^+ = s + 2$, and $g_s = \Lambda_{s+2} - \Lambda_{s+1}$. Define the sharper comparison factor $\beta = (2 - 2\tau)^{-1}$. Recover the following buffered comparison conditions in the vertex neighborhood:

$$Y_\ell - \beta X_{h^+\kappa} \geq 0, \qquad X_{hc} - \beta X_{h\kappa} \geq 0, \qquad X_{h\kappa} \geq X^{\mathrm{in}}_{h\kappa} - C_{\mathrm{fl}} N, \tag{59}$$

for all future κ, claimed c, old or newly assigned ℓ, and remaining idle h. The constant C_{fl} is fixed and sufficiently large.

For $u \geq 2$, let

$$\alpha_u = 1 + \frac{1 + 2d^{-3}}{u}, \qquad \gamma_u = 1 - \beta - \frac{\beta(1 + 2d^{-3})}{u} > 0.$$

Uniformity gives $S_\kappa \leq (1 + 2d^{-3})E/u$. The future-clock increment $c_m = \max_{\kappa \in U^+} \eta(E^m + S^m_\kappa)$ is at most $\alpha_u \eta E^m$. For a balance margin, eq. (50) gives

$$\Delta(Y_\ell - \beta X_{h^+\kappa}) \geq \gamma_u \eta E + R^Y_\ell - \beta R^X_{h^+\kappa}. \tag{60}$$

The same estimate holds for a claimed-key cap margin, with two X residuals. For sufficiently small universal E_0, the full increment is positive, which makes a recovered margin forward invariant. To retain the coefficient γ_u in the cumulative bound, we keep the residual sums explicitly. The floor margin has leading drift at least ηE.

At stage entry, the actual, unbuffered stopping margins are nonnegative. Passing to eq. (59) costs at most $C\tau(Z_s + \Lambda_H) + C\mathcal{E}_{\leq s}$ in the relevant negative parts. Indeed the difference between β and $(2 - \tau)^{-1}$ is $O(\tau)$; probability-to-odds conversion is $O(d)$ at entry, tail conversion is counted in $\chi\mathfrak{B}$, and changing future keys costs the uniformity error. The latter is also bounded by τZ_s because $Z_s \geq 3\Lambda_1/4$. Equations (57) and (58) consequently give vertex-entry deficits

$$\begin{aligned} d_0^{\mathrm{old}} &\leq (1 - \beta)D_s + \beta g_s + \text{error}, \\ d_0^{\mathrm{new}} &\leq \beta(D_s + g_s) + \text{error}, \\ d_0^{\mathrm{old\,cap}} &\leq (1 - \beta)D_s + \text{error}, \\ d_0^{\mathrm{new\,cap}} &\leq \beta D_s + \text{error}, \end{aligned} \tag{61}$$

where each error is at most $C\tau(Z_s + \Lambda_H) + C\mathcal{E}_{\leq s}$. The last line concerns the newly claimed key t: for an idle follower with entry reciprocal $X \geq Z_s - O(d)$, its cap margin after the race is $(X - 2D_s) - \beta(X - D_s)$ up to recorded errors. Its negative part is at most βD_s using $Z_s = 2D_s + O(d + \mathcal{E}_s)$. The other claimed coordinates lose only D_s. The new head starts the transition with $Y \leq 2$ and always has $Y \geq 0$, which suffices for the second line. A floor deficit is at most $D_s + C\mathcal{E}_{\leq s}$.

Until one margin recovers, sum eq. (60). If its initial deficit is d_0, the accumulated future clock is at most

$$\frac{\alpha_u}{\gamma_u}\left(d_0 + C\mathcal{E}_{\leq s} + O(\eta)\right).$$

A one-step overshoot is $O(\eta)$: this is eq. (18) for X, and follows from $|g_{\ell a_\ell}| \leq 4e_\ell$ and $Y_\ell e_\ell \leq 1 + \chi$ for Y. Recovered margins cannot recross: for $u \geq 2$ their full increment is positive throughout the cone once E_0 is small enough. Since all margins are compared against the same future clock, the time of their last recovery uses the *maximum*, not the sum, of their clock costs. A floor margin has cost at most $\alpha_u(D_s + C\mathcal{E}_{\leq s})$. Using $\alpha_u/\gamma_u = 2(u+1)/(u-1) + O(d^{-2})$ and $\beta = 1/2 + O(d^{-2})$, the total recovery clock satisfies

$$\Delta_s := \sum_{\text{sharpening}} c_m \leq \left(\frac{u+1}{u-1} + O(d^{-2})\right)(D_s + g_s) + C\tau(Z_s + \Lambda_H) + C\mathcal{E}_{\leq s}. \tag{62}$$

The additional clock for reaching $E \leq c\tau^2$ after margin recovery is at most $CHL/\tau^2 = O(d^6)$ by eq. (19); it is absorbed by τZ_s. All margins persist while this additional sharpening is performed.

If $u = 1$, necessarily $H = k = d$ and $s = d - 2$. There is one idle head and one future key κ. Write $X = X_{h\kappa}$ and $\lambda_\ell = t_\ell / p_{\ell \mathrm{BOS}} \leq \chi$. Exactly, $e_\ell = (1 + \lambda_\ell)/(Y_\ell + 1 + \lambda_\ell)$ and $p_{h\kappa} \leq 1/X$. Whenever $Y_\ell \leq (1 + \tau)\beta X$,

$$\frac{e_\ell}{p_{h\kappa}} \geq \frac{X}{Y_\ell + 1} \geq \frac{1}{(1 + \tau)\beta + X^{-1}} \geq \frac{7}{4}$$

for large d, since $X \geq X_*$. Hence $q_{\{a_\ell\}} \geq (1-\chi)e_\ell \geq (3/2)p_{h\kappa}$. Likewise $X_{hc} \leq (1+\tau)\beta X$ implies $q_{\{c\}} \geq p_{hc} \geq (3/2)p_{h\kappa}$. These estimates apply both to negative margins and to a relative neighborhood of zero. The corresponding leading drift is at least $\eta[(1-\beta)E + (3/2-\beta)p_{h\kappa}] \geq c\eta E$; the floor has the same lower bound. The simultaneous clock is at most $2\eta E$. The inventory therefore gives

$$\Delta_s \leq C(D_s + g_s + \tau\Lambda_H + \mathcal{E}_{\leq s}). \tag{63}$$

The inward argument also holds in a small relative neighborhood of a zero margin. Outside it the margin is at least $c\tau X_* \gg \eta$, so the $O(\eta)$ one-step change cannot jump back through zero. Thus these margins also persist. This exceptional case occurs at most once.

Growth of the scale at stage entry. At entry the next head's future reciprocal is $Z_s + g_s + O(\mathcal{E}_{\leq s} + d^{-3}X_*)$. Capture subtracts D_s, transition costs a recorded $O(H)$, and sharpening adds at most Δ_s plus its residual. Substituting eqs. (57) and (62) gives, for $u \geq 2$,

$$Z_{s+1} \leq \left(\frac{u}{u-1} + O(d^{-2})\right) Z_s + \left(\frac{2u}{u-1} + O(d^{-2})\right) g_s + C(d + \mathcal{E}_{\leq s} + d^{-2}\Lambda_H). \tag{64}$$

For $u = 1$, eq. (63) gives $Z_{s+1} \leq C(Z_s + g_s + \Lambda_H + \mathcal{E}_{\leq s})$. Now

$$\prod_{u=2}^{d} \left(\frac{u}{u-1} + O(d^{-2})\right) = O(d), \qquad \sum_{s<k-1} g_s \leq \Lambda_H - \Lambda_1.$$

On the stopped bound $\mathcal{E}_{\leq s} \leq 2N$, multiplying the recurrence by $d-s-2$ gives, for $s \leq d-3$,

$$(d-s-2)Z_{s+1} \leq (1+O(d^{-2}))(d-s-1)Z_s + Cd\, g_s + Cd(N + d^{-2}\Lambda_H + d).$$

Since $\sum_s g_s \leq \Lambda_H$, summation yields $Z_s \leq Cd(\Lambda_H + dN)/(d-s-1)$ for $s \leq d-2$. Summing over s and using eq. (62), with the single exceptional transition included if present, therefore gives

$$\max_s Z_s + \sum_s Z_s + \sum_{s<k-1} \Delta_s \leq Cd^2(\Lambda_H + dN) = o(\mathfrak{B}). \tag{65}$$

The sums include completed stages and the truncated current comparison stage. For a current stage whose recovery is not complete, the same inequalities bound its accumulated clock by comparing with any margin that has not yet recovered.

Cumulative error bounds. In a nonfinal cone interval, at least one future key and one idle head remain. The stopped upper band gives

$$E \geq p_{h\kappa} = p_{h\mathrm{BOS}}/X_{h\kappa} \geq (1-E_0)/(4\mathfrak{B}). \tag{66}$$

The factor 4 also covers a possible first-exit step, since $|X^+ - X| \leq 8\eta$. Throughout the sharpening interval, eq. (19) therefore gives

$$\sum_m \eta(E^m)^2 \leq CHL \sum_m \frac{E^m - E^{m+1}}{E^m} \leq CHL \log \frac{E^{\mathrm{entry}}}{E^{\mathrm{end}}} \leq CHL \log(C\mathfrak{B}). \tag{67}$$

Also $\sum_m \eta E^m \leq \Delta_s$ and $m_h \leq E$ in eq. (51). Transition changes are at most CH by eq. (18) and $Y_\ell e_\ell \leq 1+\chi$. Together with eqs. (49) and (65), these estimates give

$$\sum_s \mathcal{E}_s \leq C\theta\mathfrak{B} + Cd\chi\mathfrak{B} + Cd^3 \log(C\mathfrak{B}) + Cd^2 < N \tag{68}$$

for large d. Explicitly,

$$(\chi + d^{10}\delta_*)\mathfrak{B} = O(d^{K+18}), \qquad (d/g_*)\mathfrak{B} \leq d^{2K+7}, \qquad d\chi\mathfrak{B} = O(d^{K+19}),$$

and $(\kappa_n + \eta)\mathfrak{B} = o(1)$ by $Q \geq 8R$. The term involving g_* accounts for the follower mass when the initial scales have irregular gaps.

Preservation of the induction bounds. For any future key, eqs. (47) and (50) have common leading increments across idle heads. Therefore eq. (68) gives

$$|(X_{h\kappa} - X_{h'\kappa}) - (\Lambda_h - \Lambda_{h'})| \le CN. \tag{69}$$

Since $N = o(g_*)$, adjacent gaps exceed three quarters of their initial gaps for large d, strictly improving the stopped half-gap bound. Converting to G yields the claimed gap bound.

For two future keys write $\Delta_h = X_{h\kappa} - X_{h\kappa'}$. Equation (69) gives $|\Delta_h - \Delta_{h'}| \le CN$. Since $N = o(d^{-3}X_*)$, if $\max_h |\Delta_h| \ge \frac{1}{4}d^{-3}X_*$, all these differences have the same sign. A positive sign implies $p_{h\kappa} < p_{h\kappa'}$ for every idle head, hence $S_\kappa < S_{\kappa'}$; the sign reverses together in the other case. Thus the leading sharpening difference in eq. (50) points inward. The leading capture increments are identical across keys, and all capture, transition, and sharpening residuals total at most CN. Crossing the outer boundary in eq. (52) would require $\frac{1}{4}d^{-3}X_* \le CN + 16\eta$, which is impossible. Dividing by the reciprocal floor converts this bound into the required d^{-3} probability uniformity.

The floor is independent of any gap-to-scale ratio. Equation (53) gives $Z_s \ge 3\Lambda_1/4$ at every actual entry. During a race, a future reciprocal starts at least $Z_s - O(d^{-3}X_*)$, and eq. (57) removes at most $Z_s/2 + O(d+N)$ on every prefix. Transition costs at most CH, and vertex idle reciprocals increase exactly by Lemma 5. Consequently every future prefix remains above $\Lambda_1/8 > X_*$. The recovered comparison floor gives

$$X_{h\kappa} \ge X^{\mathrm{in}}_{h\kappa} - C_{\mathrm{fl}}N \ge (1-\tau)X^{\mathrm{in}}_{h\kappa},$$

because $N = o(\tau\Lambda_1)$.

Upper levels also improve strictly. Every nontarget idle reciprocal has at most its initial level plus twice the cumulative sharpening clock, transition changes, and recorded errors. An idle target reciprocal is bounded during its race by $G(r_h)$, whose common leading increment is nonpositive. At a later entry, a claimed idle coordinate was either a nontarget coordinate in previous races or a losing target coordinate when that key was claimed. Thus the same upper bound covers all idle raw coordinates, not just future keys. An assigned Y starts below 2 at recruitment, changes by at most $C\eta$ per step during later captures, and accumulates at most $2\Delta_s$ plus residuals in a sharpening interval. Equation (65) and eq. (68) therefore bound all these quantities and Z_s by $\mathfrak{B}$. At an actual stop, the uniformity, cap, and floor bounds imply

$$\bar{\delta} \le CHd/\Lambda_1 \le d^8\delta_*, \qquad \frac{1}{4\mathfrak{B}} \le r_h^0 \le \frac{3}{\Lambda_1}.$$

At a completed comparison recovery, $E \le c\tau^2$ and localization give $e_\ell = Y_\ell^{-1}(1 + O(\chi + e_\ell))$ and $p_{h\kappa} = X_{h\kappa}^{-1}(1 + O(m_h))$. The buffer $\beta - (2-\tau)^{-1} = \Theta(\tau)$ dominates these conversion errors. Thus eq. (59) implies all actual stopping inequalities. If the algorithm stops earlier, those inequalities hold by definition and all the pathwise bounds just proved are available at that earlier iterate. This closes the next-entry induction.

To check the remaining inequalities quantitatively, use

$$\frac{N}{g_*} \le d^{3K+30-B}, \quad \frac{N}{d^{-3}X_*} = O(d^{2K+33-B}), \quad \frac{\Lambda_H}{\Lambda_1} \le d^K,$$

and

$$(d/g_* + d^{10}\delta_* + \kappa_n + \eta)\mathfrak{B}/g_* = O(d^{3K+7-B} + d^{2K+17-B}) + o(1).$$

All tend to zero under the stated B bound. They verify probability dominance, the capture gap condition, $\Sigma = o(\tau)$, and the localization hypotheses. The strict improvements exclude every stopped boundary.

Iteration budget. Capture and transition cost $O(G(r_s)/\eta + H/\eta)$. At first vertex entry, $E^{\mathrm{entry}} \ge E_0/2$ for sufficiently small η; the total error changes by at most $CH\eta$ in a step. The upper rate in eq. (19) yields

$$(E^m)^{-2} \le 4E_0^{-2} + C\eta m, \qquad \sum_{m<T} \eta E^m \ge c(\sqrt{1+\eta T} - 1).$$

Every required recovery clock is at most $\mathfrak{B}$ by eq. (65), so recovery occurs in $O(\mathfrak{B}^2/\eta)$ steps. Sharpening to $E \le c\tau^2$ costs at most $O(HL/(\eta\tau^4)) = O(d^{10}/\eta)$ steps. Since $\mathfrak{B}^2 \le d^{2(B+K+6)}$ and $R = 4(B+K+c+30)$, the sum is less than $d^R/(2\eta)$ for large d. The final stage needs only the capture part of this argument. □

D PROOF OF THE INTERNALIZATION GUARANTEE

Proof of Theorem 1. At initialization all heads are assigned to the sink, the raw-key tolerance is zero, and total error is at most $H\delta_* < E_0$. Old-head conditions are vacuous. On the simultaneous concentration event, Lemma 10 and Lemma 9 apply successively as the row-0 target changes from $v_{A_{s-1}}$ to v_{A_s}, adding exactly one new key. Exactly one head is recruited at every stage, and all old assignments survive. The nonfinal stopping rule holds within the cap at each nonfinal stage.

After the final capture, apply Lemma 5 until $E \le d^{-(c+2)}$. This takes at most $CHLd^{2c+4}/\eta \le Cd^{2c+6}/\eta$ additional steps. Together with the capture budget in Lemma 10, this is less than d^R/η for large d. Thus the algorithm does not return FAIL.

The final assignment has one head on each secret key and all other heads on the sink. For every Boolean input, let $x_{\mathrm{BOS}} = 1$ and $e_h = 1 - p_{h,a_h}$. Then $|u_h(x) - x_{a_h}| \le 2e_h$. Since all actual and ideal head outputs have magnitude at most one,

$$|\hat{y}(x) - y(x)| \le \sum_h |u_h(x) - x_{a_h}| \le 2E \le 2d^{-(c+2)} \le d^{-c}.$$

A union bound over the concentration events for the k stages gives failure probability at most $kp = kd^{-(c+4)}$. Row decoupling makes this exactly the row-0 component of the full teacher-forced GD updates. Uniform character concentration and secret-independent entry conditions make the same sufficiently large dimension threshold valid for every fixed tuple and every set of scales satisfying eq. (7). □

Proof of Corollary 1. For independent uniform $U_h, U_{h'} \in [1, 2]$, $\Pr(|U_h - U_{h'}| < \varepsilon) \le 2\varepsilon$. Union bound over $\binom{H}{2}$ pairs gives failure at most $H(H-1)d^{-K} \le d^{-(c+5)}$. On the complement, $d^B \le \Lambda_h \le 2d^B \le d^{B+K}$ and every pair differs by at least d^{B-K}. Condition on these scales and apply Theorem 1; the sample failure bound is uniform in the realized collection. Add the two failure probabilities. □

Proof of Corollary 2. In population every nontrivial Boolean-character interaction is zero. The positive proof therefore applies deterministically with $\kappa_n = 0$, the same η, and the same caps. Every initialized raw-key vector in eq. (6) is constant, so the end-to-end initialization is key-permutation invariant. Apply Theorem 2. □

E INITIALIZATION OBSTRUCTIONS

E.1 EXACT HEAD SYMMETRY

Proof of Proposition 2. The row-0 loss is invariant under every permutation of the heads when their fixed sink scores agree. The permutation acts orthogonally on trainable logits, so its gradient is equivariant. Identical row-0 head vectors therefore receive identical gradients. Induction preserves their equality under each empirical or population GD step, irrespective of label changes between stages. Other rows do not affect this conclusion by Lemma 2.

The resulting predictor has the form

$$f(x) = \left(p_{\mathrm{BOS}} + \sum_{j=1}^{d} p_j x_j \right)^H, \qquad p_j \ge 0, \quad p_{\mathrm{BOS}} + \sum_j p_j = 1.$$

For even H, $f \ge 0$, whereas any nonempty parity is -1 on half the cube. Its squared error is therefore at least $1/2$. For odd H, f is nondecreasing in every coordinate. Fix $i \in A$ and condition on all other bits. On the event $\prod_{j \in A \setminus \{i\}} x_j = -1$, which has probability $1/2$ since $|A| \ge 2$, the labels at $x_i = -1, +1$ are respectively $+1, -1$. Writing the ordered predictions as $a \le b$, their conditional average squared error is

$$\tfrac{1}{2}[(a-1)^2 + (b+1)^2] = 1 + \tfrac{1}{2}(a^2 + b^2) + (b - a) \ge 1.$$

The other conditional pairs have nonnegative error, so the population squared error is at least $1/2$. □

E.2 SMALL GRADIENTS NEAR RANDOM INITIALIZATION

The argument starts from attention spread across many input bits with little mass on BOS. This makes head outputs typically small and yields small product gradients throughout a neighborhood of initialization.

Lemma 11 (A neighborhood with exponentially small gradients)**.** *Let $\alpha d \le H \le d$. Suppose a row-0 parameter table W^0 satisfies, for every head,*

$$p^0_{h\mathrm{BOS}} \le 1/8, \qquad \sum_{j=1}^{d} (p^0_{hj})^2 \le C_0/d, \tag{70}$$

where C_0 is a fixed constant. Then there is $\gamma > 0$ depending only on α, C_0 such that exact population GD with nonnegative step sizes and labels in $\{\pm 1\}$, even if changed between steps, remains within $\|W - W^0\|_\infty \le 1/8$ whenever $\sum_m \eta_m \le e^{\gamma d}$, for sufficiently large d. Throughout this time the squared error for the current binary target is at least $1 - Ce^{-\gamma d}$ and every raw-key probability is below $1/2$. For empirical GD using $N(d)$ independent uniform inputs drawn independently of W^0, the same statement holds except on an event of probability at most $2N(d)He^{-d/(128C_0)}$.

Proof. For fixed initialized head probabilities and uniform independent signs,

$$\mathbb{E}_x e^{\lambda \sum_j p^0_{hj} x_j} = \prod_j \cosh(\lambda p^0_{hj}) \le e^{\lambda^2 \sum_j (p^0_{hj})^2/2}.$$

Optimizing the exponential bound and using eq. (70),

$$\Pr_x \left(\left| \sum_j p^0_{hj} x_j \right| > 1/8 \right) \le 2e^{-d/(128C_0)}.$$

Let $\mathcal{G}$ be the inputs for which $|u_h(W^0, x)| \le 1/4$ for all heads. A union bound over heads gives

$$\Pr_x(\mathcal{G}^c) \le 2He^{-d/(128C_0)} =: \varepsilon_d. \tag{71}$$

For any head, $\sum_j |\partial u_h / \partial w_{hj}| = \sum_j p_{hj} |x_j - u_h| \le 2$. Hence

$$|u_h(W, x) - u_h(W^0, x)| \le 2\|w_h - w^0_h\|_\infty.$$

Throughout $\|W - W^0\|_\infty \le 1/8$, the same set $\mathcal{G}$ thus satisfies $|u_h(W, x)| \le 1/2$ for every head, and $|f_W(x)| \le 2^{-H}$. The samplewise derivative of $\frac{1}{2}(f_W - y)^2$ is

$$(f_W - y) \left(\prod_{g \ne h} u_g \right) p_{hj}(x_j - u_h).$$

It has magnitude at most $4\,2^{-(H-1)}$ on $\mathcal{G}$ and at most 4 everywhere. The population gradient therefore obeys, uniformly on the entire neighborhood,

$$\|\nabla \ell_y(W)\|_\infty \le 4(2^{-(H-1)} + \varepsilon_d) =: \beta_d = e^{-\Omega(d)}. \tag{72}$$

Choose γ smaller than half the smaller exponential rate in this bound. Before a first exit, $\|W^r - W^0\|_\infty \le \beta_d \sum_{m<r} \eta_m < 1/16$ for $\sum \eta_m \le e^{\gamma d}$ and large d. Thus a first exit is impossible in that interval. On every such iterate,

$$\mathbb{E}_x (f_W - y)^2 \ge (1 - \varepsilon_d)(1 - 2^{-H})^2 \ge 1 - \varepsilon_d - 2^{1-H} \ge 1 - Ce^{-\gamma d}.$$

Also $p^0_{hj} \le \sqrt{C_0/d}$ and a max-logit perturbation of $1/8$ changes each probability by a factor at most $e^{1/4}$. Thus every raw probability remains below $1/2$.

With $N(d)$ independent training inputs, the chance that any lies outside $\mathcal{G}$ is at most $N(d)\varepsilon_d$. On the complement every samplewise gradient has the first bound above uniformly in the neighborhood. Averaging preserves that bound, so the same first-exit argument applies to empirical GD. The bound is uniform over the neighborhood and binary labels, so the same data may be reused at every step. □

Proof of Proposition 3. For uniform logits $w^0_{hj} \in [-1, 1]$ and $b_h = 0$, deterministically,

$$p^0_{h\mathrm{BOS}} \le e/d, \qquad \sum_j (p^0_{hj})^2 = \frac{\sum_j e^{2w^0_{hj}}}{(1+\sum_j e^{w^0_{hj}})^2} \le e^4/d.$$

Thus eq. (70) holds for large d. For Gaussian logits put $V_j = e^{w^0_{hj}}$, $\mu = \mathbb{E}V_j = e^{1/2}$, and $\nu = \mathbb{E}V_j^2 = e^2$. With probability at least $1 - d^{-q}$, simultaneously over $H \le d$ heads,

$$\sum_j V_j \ge \mu d/2, \qquad \sum_j V_j^2 \le 2\nu d.$$

For each fixed integer m, expand the $2m$th centered moment of either sum. Every term with an index occurring exactly once vanishes. Surviving terms have at most m distinct indices; their moments are finite. The centered moment is therefore at most $C_m d^m$. Markov's inequality at deviations of order d, followed by a union bound over the heads, gives failure at most $C_m d^{1-m}$. Taking $m > q + 2$ proves the simultaneous assertion for large d. Consequently $p^0_{h\mathrm{BOS}} \le 2/(\mu d)$ and $\sum_j (p^0_{hj})^2 \le 8\nu/(\mu^2 d)$. Apply Lemma 11. For polynomial $N(d)$, its additional data failure is $e^{-\Omega(d)}$. The first curriculum stage uses row 0 with a singleton binary target, so failure to cross $1/2$ prevents its recruitment criterion from firing. Row decoupling prevents auxiliary-row updates from changing this conclusion. □

These independently sampled heads also have inverse-polynomial logit separation with high probability: for either distribution the difference of two logits has bounded density, so

$$\Pr\left(\min_j \min_{h \ne h'} |w^0_{hj} - w^0_{h'j}| \le \varepsilon\right) \le CdH^2\varepsilon \le Cd^3\varepsilon.$$

Thus learning can be exponentially slow even with inverse-polynomial separation of the head logits.

F End-to-end population gradient descent

Proof of Theorem 2. For a permutation $\pi \in S_d$, permute every head's raw-key coordinates and fix the sink. With $(\pi x)_j = x_{\pi^{-1}(j)}$, one has $\hat{y}_{\pi\cdot W}(\pi x) = \hat{y}_W(x)$. If $\pi A = A$, uniformity of the inputs implies $\mathcal{R}_A(\pi \cdot W) = \mathcal{R}_A(W)$, and the orthogonal coordinate action gives $\nabla\mathcal{R}_A(\pi \cdot W) = \pi \cdot \nabla\mathcal{R}_A(W)$. A key-permutation-invariant initialization is fixed by all raw-key permutations. Every exact GD iterate is therefore fixed by $S_A \times S_{A^c}$. In each head the probabilities are equal on A and equal on A^c. Global smoothness from Lemma 4 ensures a unique global gradient-flow solution, which has the same invariance.

Independently draw a key J_h from each head's attention distribution, including the sink. Let $\mathrm{Odd}(J)$ be the raw indices that occur an odd number of times. The product link and Boolean orthogonality give exactly

$$\hat{y}_W(x) = \mathbb{E}_J \chi_{\mathrm{Odd}(J)}(x), \qquad \mathbb{E}_x \hat{y}_W(x)\chi_A(x) = \Pr_J(\mathrm{Odd}(J) = A). \tag{73}$$

This event requires coverage of every key in A. Conditional on which heads land in A, their choices are independent and uniform on A. Adding further uniform draws up to a total of H can only increase coverage. The coefficient is consequently at most the probability that H independent uniform draws from A cover A.

Let Z be the number of uncovered keys in that experiment. Since $k \ge 2$ and $H/k \le 1/\alpha$,

$$\mathbb{E}Z = k(1 - 1/k)^H \ge ke^{-2H/k} \ge ke^{-2/\alpha}.$$

Changing one draw changes Z by at most one: removing one occupied key and filling one empty key, if both occur, cancel in the uncovered count. The bounded-differences inequality (McDiarmid, 1989) gives

$$\Pr(Z = 0) \le \exp\{-2(\mathbb{E}Z)^2/H\} \le \exp\{-2\alpha^2 e^{-4/\alpha} d\}.$$

Combining with eq. (73),

$$\mathbb{E}_x(\hat{y}_W - \chi_A)^2 = 1 - 2\mathbb{E}_x \hat{y}_W \chi_A + \mathbb{E}_x \hat{y}_W^2 \ge 1 - 2e^{-c_\alpha d}.$$

□

F.1 RIGHT-TO-LEFT DELETION

Consider deleting intermediate tokens in the order $z_{k-1}, z_{k-2}, \ldots, z_1$, always retaining the final answer $y = z_k$. When m intermediate tokens remain, the output sequence is $z_1, \ldots, z_m, y$. Keep the absolute parameter rows and fixed sink scores of Section 3, with no copying of parameters between rows. Row $j < m$ predicts z_{j+1} from $[\mathrm{BOS}, x_1, \ldots, x_d, z_1, \ldots, z_j]$, and row m predicts y from the same form of key list. Consequently, row 0 always sees only the raw inputs and sink, and its target is

$$\begin{cases} z_1 = x_{t_1}, & m \geq 1, \\ y = \prod_{i=1}^{k} x_{t_i}, & m = 0. \end{cases}$$

By the row-decoupling argument of Lemma 2, every deletion before the last leaves this row's objective unchanged. With matched batches and step sizes, its empirical trajectory up to the final deletion is exactly continued training on z_1. The answer moves to a different parameter row at each deletion, so assignments learned in the preceding answer row are not transferred to the next one.

Proof of Corollary 3. Let $A = \{t_1, \ldots, t_k\}$, $A' = A \setminus \{t_1\}$, and $q = |A'| = k - 1$. Both possible row-0 targets, x_{t_1} and χ_A, are invariant under permutations of A'. The population loss is invariant and its gradient is equivariant under these permutations, as in the proof of Theorem 2. The initialization is fixed by them, so every update preserves equality of the raw-key logits within A' in each head. Changing the target at the final deletion preserves the same symmetry. Uniqueness gives the analogous statement for gradient flow.

The coefficient identity eq. (73) applies to the resulting row-0 predictor. The event $\mathrm{Odd}(J) = A$ requires every key in A' to be drawn. Conditional on the heads landing in A', their choices are independent and uniform on A'. Its probability is therefore at most the coverage probability of H independent uniform draws from q keys. If Z counts uncovered keys, then $q \geq 2$, $q \geq \alpha d$, and $H \leq d$ give

$$\mathbb{E}Z = q(1 - 1/q)^H \geq qe^{-2H/q} \geq \alpha d e^{-2/\alpha}, \qquad \Pr(Z = 0) \leq e^{-2(\mathbb{E}Z)^2/H} \leq e^{-c_\alpha d}.$$

The probability bound uses the same bounded-differences estimate as Theorem 2. Hence $\mathbb{E}_x(\hat{y}_{W^r} - \chi_A)^2 = 1 - 2\mathbb{E}_x \hat{y}_{W^r} \chi_A + \mathbb{E}_x \hat{y}^2_{W^r} \geq 1 - 2e^{-c_\alpha d}$. □

Under these population dynamics, one-bit pretraining can distinguish t_1 but leaves $t_2, \ldots, t_k$ indistinguishable within each head. For $k = 2$, the two deletion orders coincide.

G EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

G.1 SETUP

For Table 2, we use $d \in \{32, 64\}$, $k \in \{8, 12, 16\}$, and $H = k$. The target is the parity of the first k input coordinates. Each setting uses ten independent training seeds. The model, initialization, optimizer, and stage budget are the same as in Figure 2. Final-answer accuracy is measured on 100,000 independent uniform inputs shared among the three methods within each seed.

Initialization. In every head and row, draw independent preliminary logits from $\mathcal{N}(0, 1.5^2)$ at all active keys, then draw an independent sink probability $q_h \sim \mathrm{Unif}[0.8, 0.9]$. Retain the Gaussian non-sink logits and replace the sink logit by

$$w_{h,\mathrm{BOS}} = \log\left(\sum_{i \neq \mathrm{BOS}} e^{w_{h,i}}\right) + \log\frac{q_h}{1 - q_h}.$$

Calibration after the noise gives actual sink attention q_h, so every head initially favors the sink. All logits, including the sink, remain trainable.

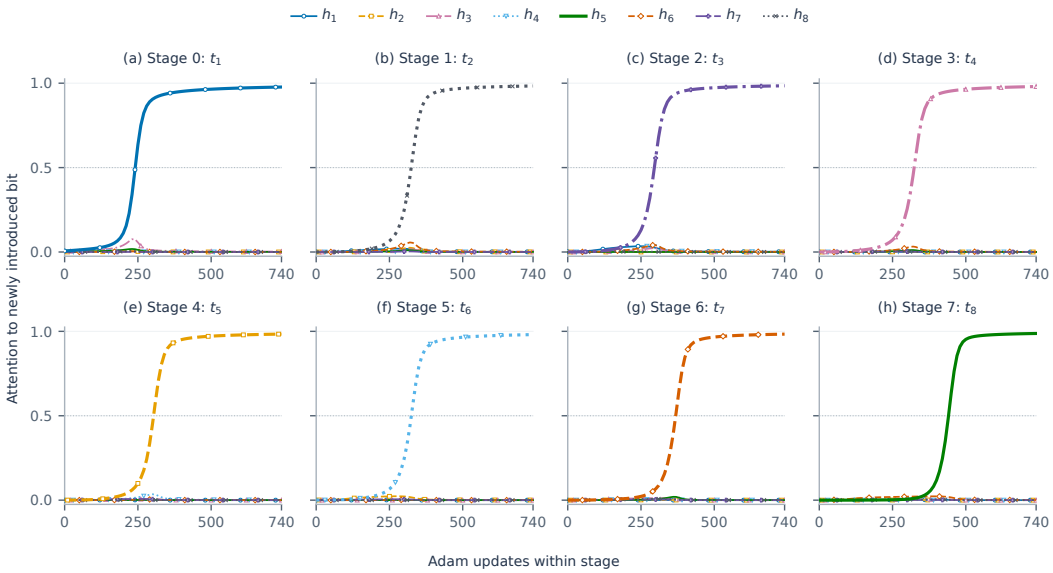


Figure 3: One head wins the competition for each newly required bit throughout the curriculum in Figure 2. Panels show all eight heads' attention at successive stages; thicker curves identify the recruited heads.

G.2 Sequential head recruitment

Figure 2 uses the $d = 32$, $k = H = 8$, seed-0 run included in Table 2. This setting and seed were fixed before the grid experiment. In this displayed run, every active row has full-pool half-MSE $\leq 10^{-2}$ and zero sign errors at every endpoint. All row-0 endpoints pass the all-input sign certificate; both CoT methods achieve 100% final test accuracy. Figure 2(a) plots the average per-token loss,

$$\frac{1}{2|B|R} \sum_{x \in B} \sum_{j=0}^{R-1} \left(\hat{z}^{(j)}(x) - z^{(j)}(x)\right)^2,$$

where $z^{(j)}$ is row j's target, B is the first $1{,}024$ training inputs, and R counts active tokens: 1 for Direct, 8 for CoT, and $8 - s$ for internalization at stage s. The training objective remains summed per-row MSE. Matched initial parameters, targets, and updates give identical first-stage CoT curves; subsequently, deletions increase internalization's loss while CoT continues the unchanged task.

Figure 2(b) labels maximum-attention keys in fixed head order. Green shading gives the selected key's attention weight on a common $[0.7, 1]$ scale, including sink cells, labeled BOS. Bold keys and thick outlines mark newly recruited heads; thin outlines mark retained assignments. All heads initially select the sink and eventually specialize to distinct target bits.

Figure 3 plots all eight heads separately throughout each 740-update stage. At stage s, row 0 predicts the parity of $t_1, \ldots, t_{s+1}$, and the panel shows attention to the newly introduced bit t_{s+1}. Row 0 attends only to BOS and raw inputs in every stage. One head wins each stage's competition, while earlier assignments persist throughout subsequent stages.

G.3 Head similarity and specialization

We isolate the first recruitment stage of the $d = 32$, $k = H = 8$ model: all eight heads attend to BOS and the raw inputs, and their product predicts x_1. Holding the target fixed avoids effects from previously recruited heads or earlier stage durations. This is a single-stage ablation, not an additional full-curriculum comparison.

Controlling head similarity. For each of ten seeds, draw an 8×32 matrix Z of independent standard Gaussians. Subtract each column's mean across heads and divide the centered matrix by its

root mean square, obtaining $\widetilde{Z}$. Set

$$w_{h,i} = \sigma \widetilde{Z}_{h,i}, \qquad w_{h,\mathrm{BOS}} = \log\left(\sum_{i=1}^{32} e^{w_{h,i}}\right) + \log\frac{0.85}{0.15}.$$

Thus σ is exactly the root-mean-square cross-head raw-logit deviation, and every head starts with sink attention 0.85 at every spread. The conditions are $\sigma \in \{0, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.5\}$; zero gives exactly identical heads. Within each seed, all seven conditions share $\widetilde{Z}$, the training pool, minibatches, and evaluation inputs.

Measuring specialization and learning. We retain the main experiment's objective, Adam settings, training-pool size, and minibatch size. Each condition receives $30{,}000$ updates without optimizer resets or early stopping; this budget was chosen using a separate pilot.

We measure attention and half-MSE on a fixed independent $4{,}096$-input probe every ten updates. Let $p_{(1)}(t) \geq p_{(2)}(t)$ be the largest two heads' attention weights on x_1. Separation time is the first recorded update with $p_{(1)}(t) - p_{(2)}(t) \geq 0.25$; learning time is the first with probe half-MSE $\leq 10^{-2}$. These criteria measure visible head separation and prediction accuracy separately. Unreached thresholds are censored at $30{,}000$. Final accuracy uses $100{,}000$ fresh uniform inputs per seed.

Table 3: Nearly identical heads take longer to specialize, but all nonidentical conditions learn the first-stage target. Exactly identical heads do not learn within the budget. Times are median [IQR] updates and accuracy is mean test accuracy (%), over ten paired seeds.

Spread σ	Separation	Half-MSE $\leq 10^{-2}$	Learned	Test acc.
0	not reached	not reached	0/10	50.05
10^{-6}	9,475 [9,200, 9,830]	9,510 [9,232, 9,862]	10/10	100.00
10^{-4}	3,185 [3,015, 3,440]	3,250 [3,070, 3,502]	10/10	100.00
10^{-3}	1,215 [1,140, 1,322]	1,300 [1,220, 1,412]	10/10	100.00
10^{-2}	460 [435, 478]	555 [532, 588]	10/10	100.00
10^{-1}	260 [250, 270]	380 [372, 380]	10/10	100.00
1.5	175 [155, 188]	320 [320, 338]	10/10	100.00

Learning after head separation. The individual-head trajectories use the preselected seed 0. Figure 4 and table 3 show that smaller perturbations mainly prolong the interval before one head separates. Median separation times are $9{,}475$ and 175 updates for $\sigma = 10^{-6}$ and 1.5, respectively; median learning times are $9{,}510$ and 320. The median additional time after separation is only 30 updates in the near-identical condition, compared with 150 for the larger spread. The near-identical condition is slower in every paired seed. The comparison also holds with separation thresholds 0.1 or 0.4, or loss thresholds 0.02 or 0.005.

G.4 Curriculum failure and recovery

Section G.3 shows that nearly identical heads take longer to specialize. Here we illustrate how incomplete learning can interact with a fixed deletion schedule. We repeat the main $d = 32$, $k = H = 8$ curriculum with ten paired seeds, contracting each head's non-BOS logits and initial BOS attention toward their across-head means by a factor ρ, then recalibrating the BOS logit. This preserves the initial BOS preference; $\rho = 1$ retains the main initialization and $\rho = 0$ makes heads identical. We compare 740 and $1{,}480$ updates per stage, keeping the remaining training settings unchanged.

Among the 70 runs with nonidentical initial heads, only one fails at the shorter budget: seed 3 at $\rho = 10^{-4}$. We examine this failure and its matched longer-budget run. With 740 updates per stage, some prefixes are learned but later stages remain underfit, and several heads favor the same bits (Figure 5a,b). Doubling the stage budget resolves these repeated assignments into one head per target bit (Figure 5c), improving final accuracy from 50.1% to 100%. The recovered run still underfits its first two prefixes, so an undertrained stage need not imply final failure. All other nonidentical initializations succeed at both budgets; this is an isolated failure case, not a monotone relationship between head similarity and final success.

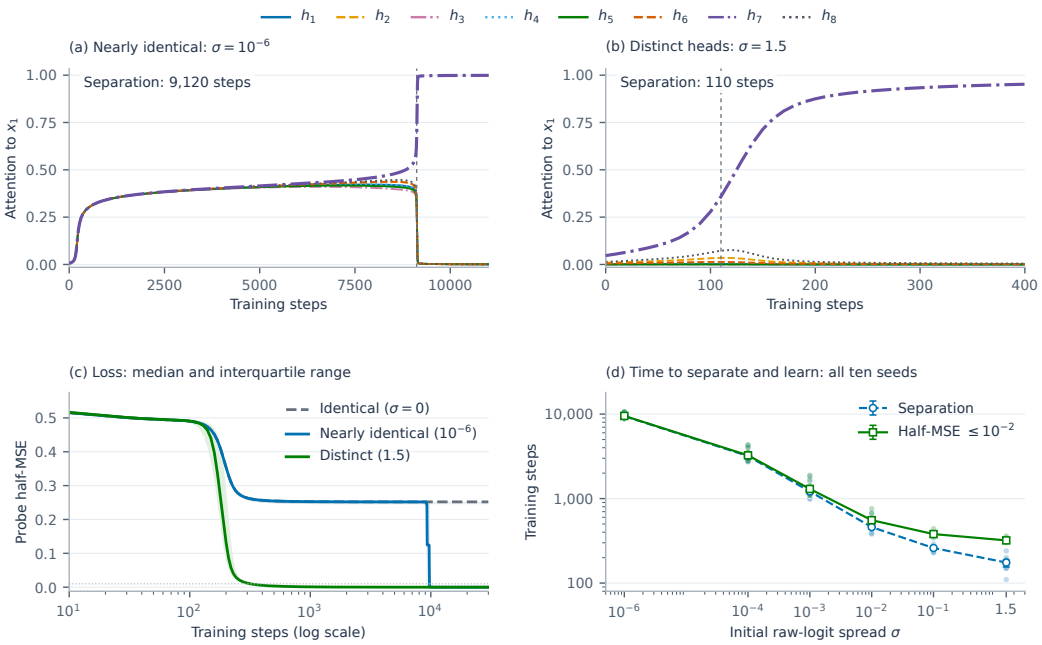


Figure 4: Nearly identical heads take longer to specialize. (a,b) Individual-head races illustrate the difference. (c,d) Most of the extra training occurs before separation; learning follows soon afterward. Summaries show medians and interquartile ranges over ten seeds.

The example shows that a fixed deletion schedule can advance before the current stage is learned, while additional training can enable recovery. Together with the slower specialization observed in Section G.3, this offers a possible mechanism for the token-removal schedule sensitivity reported by Deng et al. (2024).

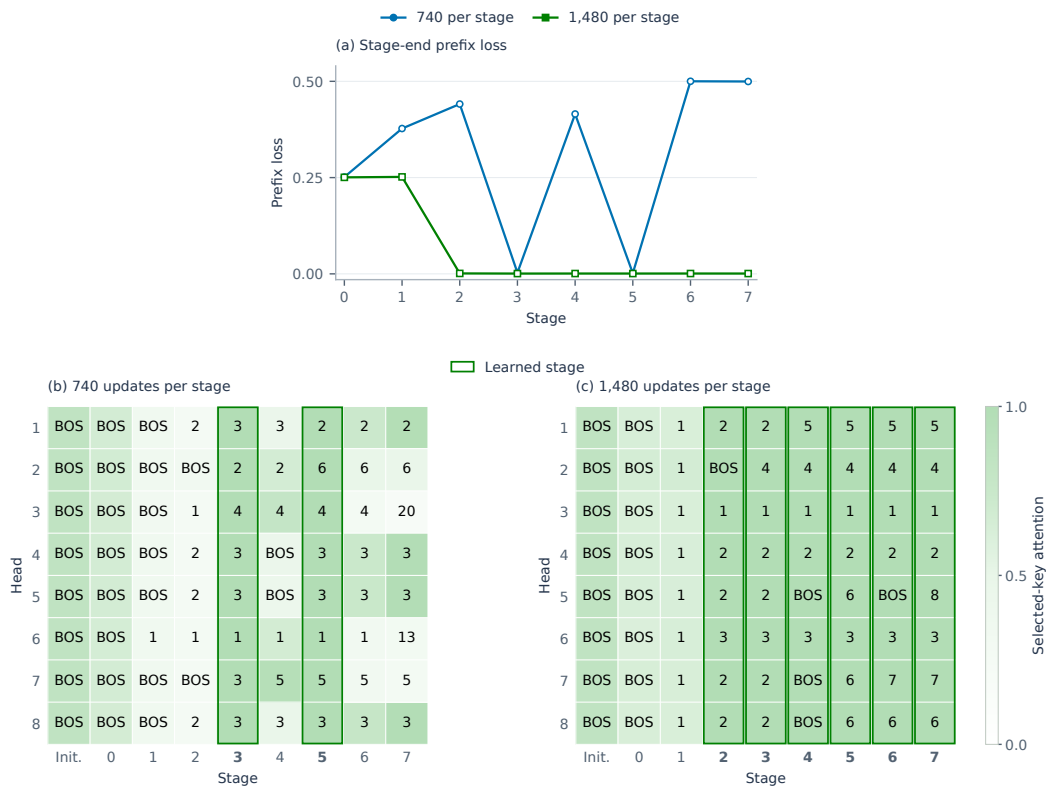


Figure 5: More training per stage rescues an isolated internalization failure. (a) Stage-end prefix losses for the failed run and its longer-budget pair. (b,c) Head assignments, with green outlines marking successfully learned stage columns. Shading indicates the selected key's attention weight.

G.5 Internalization without an initial BOS preference

We repeat the complete grid of Table 2, changing only the initial BOS logits. Every active logit in every head and parameter row, including BOS and available prefix-token keys, is drawn independently from $N(0, \sigma^2)$ with $\sigma = 1.5$.

Comparison with the main experiment. The grid uses $d \in \{32, 64\}$ and $k = H \in \{8, 12, 16\}$, with ten paired seeds per setting and the same 740 updates per stage as the main experiment.

Table 4: Removing the initial BOS preference reduces reliability relative to Table 2 at the same budget; runs finish near perfect accuracy or chance. Final-answer accuracy (%), mean ± SD over ten seeds, with success counts below.

Method	CoT tokens	Stages	$d = 32$			$d = 64$		
			$k = 8$	$k = 12$	$k = 16$	$k = 8$	$k = 12$	$k = 16$
Direct	0	1	50.0 ± 0.2	50.0 ± 0.1	50.0 ± 0.1	50.0 ± 0.2	50.0 ± 0.1	49.9 ± 0.1
Explicit CoT	$k-1$	1	100.0	75.0 ± 26.4	75.0 ± 26.3	65.1 ± 24.1	49.9 ± 0.1	50.0 ± 0.1
CoT + internalization	0	k	90.0 ± 21.1	80.0 ± 25.8	80.0 ± 25.9	90.0 ± 21.1	50.0 ± 0.1	50.0 ± 0.2
Successes ($\geq 99\%$ accuracy)			Number of seeds out of ten					
Direct			0/10	0/10	0/10	0/10	0/10	0/10
Explicit CoT			10/10	5/10	5/10	3/10	0/10	0/10
CoT + internalization			8/10	6/10	6/10	8/10	0/10	0/10

Accuracy and recovery with longer training. Table 4 summarizes final-answer accuracy, with success defined as at least 99%. Internalization succeeds in $(8, 6, 6)/10$ seeds at $d = 32$ and $(8, 0, 0)/10$ at $d = 64$, for $k = 8, 12, 16$, respectively. Explicit CoT succeeds in $(10, 5, 5)/10$ and $(3, 0, 0)/10$; Direct remains near chance throughout. Every successful run has 100% test accuracy, and all successful internalization predictors pass the all-input sign certificate. Thus the success counts also equal the zero-test-error counts. The large standard deviations reflect outcomes near either 50% or 100%, rather than consistently intermediate accuracy. Ten-seed estimates remain imprecise: the 95% Wilson intervals for $0/10$, $6/10$, $8/10$, and $10/10$ are approximately $[0, 0.28]$, $[0.31, 0.83]$, $[0.49, 0.94]$, and $[0.72, 1]$.

An exploratory follow-up at $d = 32$, $k = H = 8$ doubles every stage to $1{,}480$ updates for the same ten seeds and all three methods. Internalization improves from $8/10$ to $10/10$ successes; CoT remains $10/10$ and Direct $0/10$. Both failed runs recover from chance to perfect accuracy, and every stage then meets the row-0 learning criteria. This establishes budget sensitivity in this setting.

Pair cancellation and head specialization. We fixed $d = 32$, $k = H = 8$, seed 0 as the illustrative run before training. Figure 6 follows its losses and head assignments. At the first endpoint, the selected keys are $(1, 29, 30, \mathrm{BOS}, 24, 24, 29, 30)$. The repeated raw keys cancel because $x_i^2 = 1$, leaving the correct one-bit parity. Thus assignments outside the target support need not imply failure. A cancelling pair can later separate onto distinct target bits as the prefix grows.

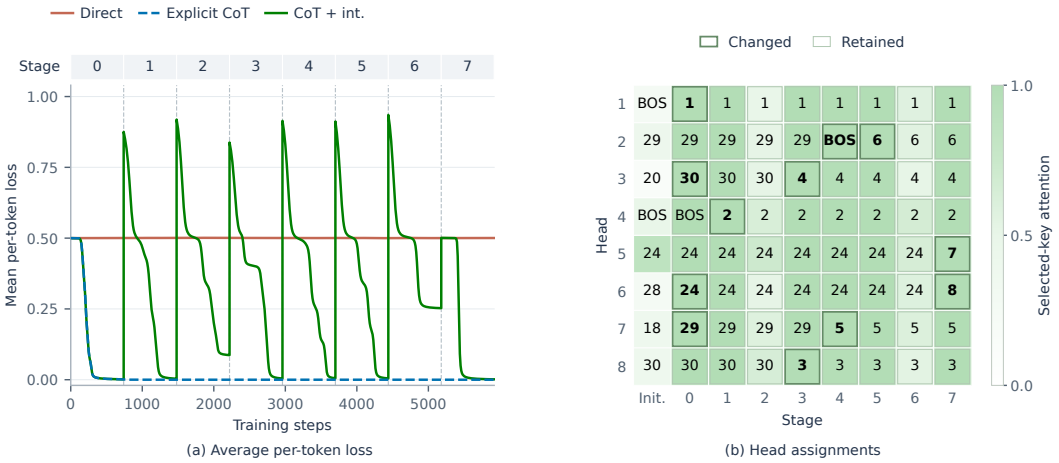


Figure 6: Internalization can succeed through cancelling head pairs without an initial BOS preference ($d = 32$, $k = H = 8$). (a) Some stages remain underfit despite eventual success. (b) Heads form cancelling pairs before separating onto distinct target bits.

The illustrated run leaves stages 2 and 6 underfit, but recovers in later stages and ultimately assigns one head to each target bit, achieving 100% accuracy. Earlier learned assignments need not persist throughout this process. Figure 7 shows every head separately, including the stages in which the newly introduced bit is not captured within the budget.

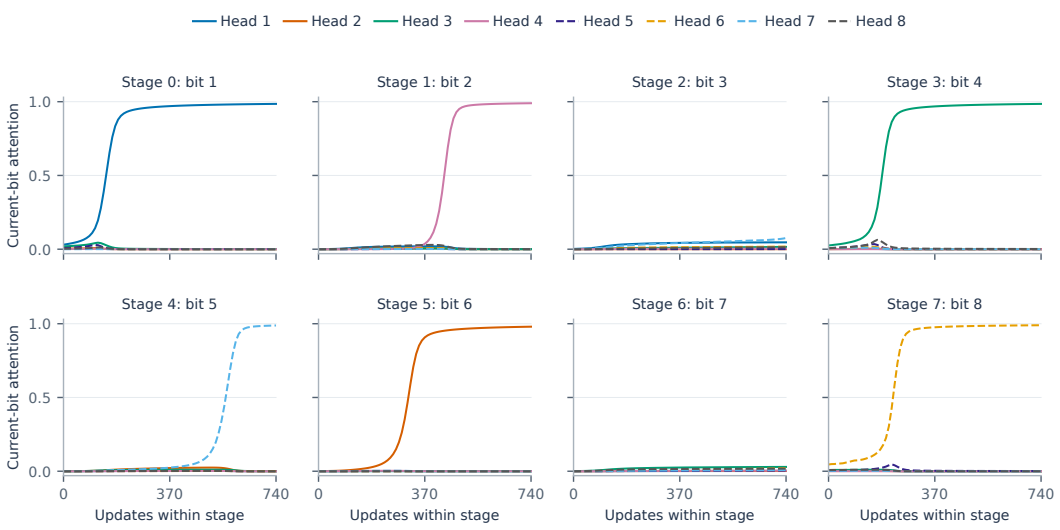


Figure 7: New bits need not be captured at every stage. Individual-head trajectories for Figure 6 show no head capturing the new bit at stages 2 and 6, followed by recovery in later stages.

G.6 Internalization with extra heads

We repeat Table 2 at $d \in \{32, 64\}$ and $k \in \{8, 12, 16\}$ with $H = k + 1$ and $H = k + 2$, using ten seeds per setting. The initialization, optimizer, one-bit starting prefix, and k-stage curriculum follow the main protocol, with the same training budget of 740 updates per stage. All three methods are retrained at each head count.

Table 5: Internalization remains effective with extra heads. Final-answer accuracy (%), mean ± SD over ten seeds.

H	Method	$d = 32$			$d = 64$		
		$k = 8$	$k = 12$	$k = 16$	$k = 8$	$k = 12$	$k = 16$
$k + 1$	Direct	50.0 ± 0.2	50.0 ± 0.2	50.0 ± 0.1	50.1 ± 0.1	50.0 ± 0.1	50.0 ± 0.2
	Explicit CoT	100.0	100.0	100.0	100.0	100.0	100.0
	CoT + internalization	100.0	100.0	100.0	100.0	100.0	100.0
$k + 2$	Direct	50.0 ± 0.1	50.0 ± 0.1	49.9 ± 0.2	49.9 ± 0.1	50.0 ± 0.1	50.0 ± 0.2
	Explicit CoT	100.0	100.0	100.0	100.0	100.0	100.0
	CoT + internalization	100.0	100.0	100.0	100.0	100.0	100.0

Explicit CoT and internalization attain perfect final-answer accuracy in every run reported in Table 5, while Direct remains near chance. Every successful internalization run assigns one head to each target bit and the surplus heads to BOS.

Figure 8 illustrates recruitment at $d = 32$, $k = 8$ for seed 0. Each stage recruits a head for the next bit while preserving earlier assignments; the one or two unused heads remain on BOS. Successful internalization does not require the number of heads to match the parity size exactly: surplus heads can approximate constant factors while the others encode the target parity.



Figure 8: Sequential recruitment with extra heads ($d = 32$, $k = 8$). Target assignments persist, while one or two surplus heads remain on BOS.