LEARNING TO EXECUTE TREE SEARCH OVER CONTINUOUS THOUGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tree-structured reasoning lets a language model explore several intermediate thoughts, but the branching, scoring and backtracking are usually controlled by an algorithm outside the model. We ask whether a Transformer can execute that search internally, as one autoregressive chain of thought over the continuous thoughts it generates and feeds back into its own context, and whether standard training delivers this ability. We study search on depth-H trees with the search history represented as a sequence of vector-valued tokens in the Transformer's context. The Transformer predicts the next token to select, backtrack or terminate, while a fixed front end supplies candidates with noisy rewards. For the search itself, we construct a binary instance in which single-path success has an upper bound that decreases polynomially with depth, while depth-first search succeeds with probability $1-H^{-3}$, so backtracking is essential to the search we execute. On the representation side, we construct a Transformer with one attention layer that executes depth-first search exactly by predicting successive tokens, up to the first reward gap below a threshold, for every fixed branching factor. On the learning side, we give two answers to whether a Transformer trained by empirical risk minimization (ERM) executes it. First, because the correct next action is determined by the prefix, on-policy imitation learning with ERM learns a single executor without quantizing thoughts, with failure $\widetilde{O}(D/N)$ for D proportional to the trainable parameter count and N independent instances per round. However, this approach requires repeated collection and labeling of the learner's own rollouts. Second, under teacher forcing with a quantized front end, the guarantee depends on what is written back. With raw feedback, our guarantee can require exponentially small training errors as the horizon grows. Our counterexample class contains near-optimal models that leave the teacher's trace. Writing the quantized state back restores a constant tolerance, under which ERM over the executor's sparse template reaches the goal within $5H$ selections with probability $1 - H^{-3} - \widetilde{O}(H/N)$. Together, our results show that a Transformer can execute search over its own continuous thoughts and clarify how guarantees of learned execution differ between on-policy imitation learning and teacher forcing, and depend on how predictions are fed back into context.

1 INTRODUCTION

Language models can improve their reasoning by using additional test-time computation to explore and evaluate candidate solutions (Snell et al., 2025). Tree of Thoughts and related methods (Yao et al., 2023; Hao et al., 2023; Zhou et al., 2024; Besta et al., 2024) improve reasoning performance by generating and evaluating candidate thoughts and exploring alternative reasoning paths. However, the search itself is controlled by an algorithm outside the language model.

A chain of thought (Wei et al., 2022) already carries a state through the model's own context without an external controller, but a chain restricted to parent-to-child moves follows a single path. This raises the following question:

Can a Transformer execute tree search including backtracking without an external search controller, using a single autoregressive chain of thought over the continuous thoughts it generates? Can standard training deliver such an executor?

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.

Answering these questions would connect learning from search traces to guarantees of reliable search at inference time. Coconut (Hao et al., 2025) feeds continuous hidden states back into the model and observes search-like behavior. However, the same feedback can propagate prediction errors to later steps, making it nontrivial to guarantee correct search execution. Even if a Transformer can execute the search, training does not necessarily produce such an executor. Forward demonstrations can fail to teach lookahead (Bachmann & Nagarajan, 2024; Saparov et al., 2025), whereas traces containing exploration and backtracking can improve learning (Gandhi et al., 2024; Yang et al., 2026b; Kim et al., 2025; Alberghi et al., 2025). These findings motivate studying which training conditions are sufficient for correct search execution at inference time.

We study this question in a family of noisy depth-H tree-search environments. The search history is represented in the Transformer's context as a sequence of vector-valued tokens carrying continuous thoughts and discrete search information. The Transformer advances the search by predicting the next token to select, return or terminate. A fixed front end supplies candidate thoughts from the current thought, one per child, each with a node embedding and a reward. One may read the front end as a frozen pretrained proposer paired with a reward signal from the environment; the setting then corresponds to post-training for search control with the proposer fixed, and the question is whether search control is learned. We first compare single-path procedures with backtracking search in our noisy tree environments. We then construct a Transformer that executes the search over continuous thoughts. Finally, we derive learning guarantees for supervision on the model's own histories and on fixed teacher traces. We also test whether standard optimization learns search and how feedback storage affects execution.

Contributions.

- **(C1) Backtracking enables reliable search** (Theorem 1). We show that, on our noisy binary-tree family, single-path success vanishes with depth, while depth-first search succeeds with high probability using linearly many steps.
- **(C2) Exact execution with one attention layer** (Theorem 2). We construct a Transformer with one attention layer that exactly executes reward-ordered depth-first search over continuous thoughts by autoregressive token prediction, including backtracking, under a reward-gap condition.
- **(C3) Learning through on-policy imitation learning** (Theorem 3). We show that ERM on accumulated histories from the model's own rollouts learns an executor without thought quantization. Its finite-sample closed-loop guarantee holds until the reward-gap condition fails or the search stops.
- **(C4) Learning through teacher forcing** (Theorems 4 and 5). We give raw-feedback execution guarantees when training errors are sufficiently small, but show that near-optimal expected loss alone need not ensure path success in the counterexample class of Setting 4. Full write-back gives path and goal guarantees for the analyzed executor classes with constant training-error tolerance.

Our results show that a Transformer can execute backtracking search over its own continuous thoughts and establish conditions under which such execution can be learned.

Related work. Several recent works construct or train Transformers that execute search. Kim et al. (2026) pose unknown tree search with bandit feedback, where tree extensions and feedback are supplied externally, and show that Transformers can express distinct search strategies and approximate them when trained on traces. Yang et al. (2026a) prove that policy-gradient training in a k-ary tree with a terminal reward makes a two-head Transformer learn randomized depth-first search with backtracking. Lehnert et al. (2024) train Searchformer on token traces of A* search dynamics, and Back de Luca & Fountoulakis (2024) simulate depth-first search and shortest paths on explicit graph inputs with looped Transformers. These works study search over discrete nodes or symbolic states, rather than model-generated continuous thoughts. Kim et al. (2026) and Lehnert et al. (2024) evaluate learning empirically without theoretical learning guarantees, while Back de Luca & Fountoulakis (2024) construct executors but do not analyze how they can be learned. Zhu et al. (2025) construct a Transformer that uses continuous thoughts to solve reachability on explicitly given graphs, and Zhu et al. (2026) analyze how this mechanism is learned. We instead study search over continuous thoughts, establishing both exact execution and finite-sample learning guarantees.

Our learning guarantees belong to the theory of imitation learning. Ross et al. (2011) established imitation-learning guarantees for DAgger, which aggregates expert-labeled data from the learner's

own rollouts. We build on this approach to prove that Transformers can learn to execute search over their own continuous thoughts (Theorem 3). Simchowitz et al. (2025) show that imitation errors can grow exponentially during execution in continuous-action control. We study this issue in Transformer search over continuous thoughts, giving sufficient conditions for reproducing the teacher's search path and showing that nearly optimal teacher-forced risk can still lead to path failure in the counterexample class of Setting 4 (Theorem 4). Prior work shows that discrete-token CoT expands Transformers' computational power (Merrill & Sabharwal, 2024; Feng et al., 2023; Li et al., 2024), and Kim & Suzuki (2025) establish efficient learning guarantees for parity with CoT. Full write-back similarly restricts stored thoughts to a finite set. We compare its tree-search execution guarantees with those under raw continuous feedback (Theorems 4–5). Empirically, Phua et al. (2026) and Patel (2026) report Transformers that fit backtracking traces yet fail to execute the search; Theorem 4(b) is a counterpart inside the class, a member near the optimal teacher-forced risk that leaves the teacher's trace.

2 Setting

2.1 Autoregressive search representation

We represent search histories as vector-valued token sequences in the Transformer's context. Selection, backtracking and termination are expressed by predicting the next token. We call the token format and its execution rules the thought interface. Figure 1 illustrates how selection and backtracking are represented in the token sequence.

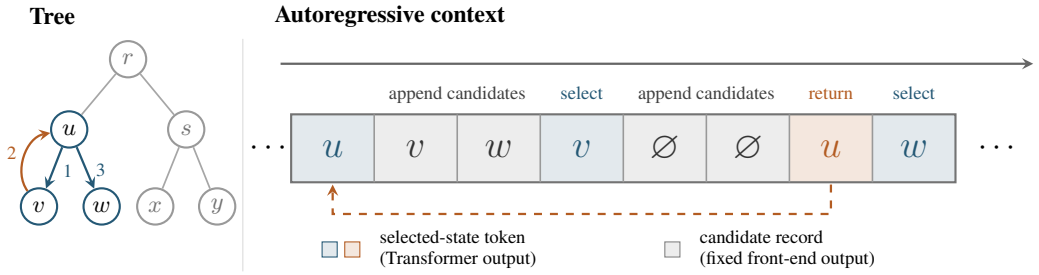


Figure 1: **Search in a token sequence.** Each cell represents one vector-valued token; only its node or validity is shown. The Transformer selects v, returns to u, then selects w. Here, v is non-accepting with no valid children and the budget permits continuation. The return appends a token with updated bookkeeping; gray cells are candidate records.

Node embeddings, thoughts and the front end. We consider a tree of depth H and branching factor $C \geq 2$, with $S = (C^{H+1} - 1)/(C - 1)$ nodes. We use fixed *node embeddings* to identify nodes and continuous thoughts to represent search states. We choose embeddings $a_u \in \{\pm 1/\sqrt{d}\}^d$ with $d = \Theta(\log S)$ that are pairwise incoherent: $|\langle a_u, a_v \rangle| \leq \varepsilon_{\text{emb}}$ for $u \neq v$ (Lemma A.1); thoughts satisfy $z \in \mathbb{R}^d$ and $\|z\|_2 \leq Z$.

Candidate slots at each node are indexed by $b \in [C] := \{0, \ldots, C-1\}$. For slot one-hot e_b, the fixed *front end* returns a candidate thought, child embedding and reward: $g_{\text{pre}} : (z, a_u, e_b) \mapsto \big(z_b(z,u), a_{v_b(z,u)}, r_b(z,u)\big)$. An invalid slot returns a_{eos}, the end-of-sequence embedding. The Transformer controls selection, return and termination, but does not generate candidates. The node embeddings and front end satisfy the following assumption.

Assumption 1 (Bounded interface)**.** Node embeddings and reserved embeddings a_0 (absent parent) and a_{eos} (end of sequence) have pairwise incoherence $|\langle a, a' \rangle| \leq \varepsilon_{\text{emb}}$ for distinct embeddings a, a', with $0 < \varepsilon_{\text{emb}} < 1/2$. For every thought of norm at most Z, g_{pre} returns candidate thoughts of norm at most Z and rewards of magnitude at most R. Valid child embeddings are distinct members of the fixed embedding set other than a_0 and a_{eos}. This assumption requires no Lipschitz bound or contraction, and generated thoughts need not be near the teacher's.

Tokens, records, and anchors. We call the token sequence a trace. A selected-state token represents a node visit: its coordinate blocks hold the thought z, a source thought z_{src} retained for returns, the

current embedding U and the parent embedding P, together with candidate-embedding and reward blocks V, r, scratch coordinates, and bookkeeping such as the mask of exhausted slots and a backlink to the parent visit (Definition E.1). The search structure is thus encoded in the token sequence, without a separate tree or stack for storing the search history. A selected-state token represents either a new node visit or a return to an earlier visit. Each child selection starts a new visit, even if its node embedding has appeared before. If execution continues after the token starting a new visit, the C *candidate records* returned by g_{pre} from that thought are appended to the same context. The Transformer is queried only at *anchors*, the columns at which a control decision is due: the last candidate record of a new visit, or a return token.

The search rule. We define the *search rule* as reward-ordered DFS on the tree generated by the front end (Definition E.2). At each anchor before stopping, the search rule determines the next output from the recorded history. A selection chooses the largest-reward valid candidate whose slot is not completed and copies its thought into both thought blocks of a token for the new visit. When none remains, the rule returns to the linked parent visit and marks the returning child's slot as completed, or terminates at an exhausted root. Applying the rule from the root defines the *teacher trace*. A fixed rule specifies which decoded nodes count as goals; in the environment of Section 3, only the target leaf is a goal. The budget B_{sel} limits the number of child selections. Execution stops at acceptance, termination or B_{sel} child selections (Definition E.3). Counting selections, returns and termination, execution requires at most $B := 2B_{\mathrm{sel}} + 1$ next-token predictions. Each prediction is checked before any acceptance, termination or budget stop that it triggers.

To carry out discrete operations, the interface decodes node-embedding blocks to their nearest predefined embeddings and other discrete blocks to their nearest predefined values. These values determine action dispatch, the front-end address, acceptance, termination and budget stopping. By default, we use *raw feedback*: predictions remain in context after clipping only the two thoughts to radius Z, and decoded values are read only to act.

Success notions. We define the *decoded path* as the sequence of selections and returns, including their discrete blocks and termination. *Path success* means that the learner's decoded path equals the teacher's; *goal success* means that a selected decoded node is accepting. At inference time, however, the model acts on its own thoughts and may encounter candidates that differ from the teacher's. *Closed-loop correctness* instead requires the decoded discrete outputs at each anchor before stopping to match the search rule applied to the decoded view of the rollout's own raw prefix. It evaluates decisions on the model's generated tree, not against a teacher trajectory. Our execution and on-policy guarantees use reward gaps to delimit their scope. At a selection, the *eligible reward gap* is the difference between the two largest rewards among valid uncompleted candidates ($+\infty$ for a singleton). For a threshold t, we define τ_t as the first prediction step at which a due selection has a reward gap below t.

2.2 Transformer class and learning protocol

Architecture. We use norm-bounded causal Transformers with positional embeddings concatenated with token features and no layer normalization (Definition A.2). For token matrix $X = [x_1, \ldots, x_m]$, an attention layer and its following FFN compute $y_i = x_i + \sum_{h=1}^{J} V_h X_{:,\le i} \operatorname{softmax}\big(X_{:,\le i}^{\top} B_h x_i\big)$ and $x_i^{+} = \psi(y_i)$. Here, J is the number of heads, B_h and V_h are the score and value matrices, and ψ is a columnwise ReLU FFN. Attention reads only columns up to i; the FFN replaces its input without a residual connection. The class consists of an initial FFN followed by these attention/FFN pairs. Width is the token dimension and context the number of stored columns. The *score radius* bounds the score matrices; the *FFN radius* bounds FFN weights and biases. The main-text asymptotic bounds assume fixed attention and FFN depths, fixed $(C, \varepsilon_{\mathrm{emb}}, Z, R)$, and parameter norm bounds and covering-number constants polynomial in the training sample size, H and $1/\delta$; constants may depend on C, $\varepsilon_{\mathrm{emb}}$, Z and R (Appendix A.1).

Training. We compare training on learner-generated histories with training on fixed teacher traces. Both protocols use empirical risk minimization (ERM) in real arithmetic. We write $N \ge 1$ for the number of independent instances per round in on-policy imitation learning, or in the fixed dataset for teacher forcing. In both protocols, $F(E)_{:,-1}$ is the last-column prediction on the predicted coordinates. Probabilities over independent test instances are conditional on the training history.

In *on-policy imitation learning*, we roll out the current Transformer on N new i.i.d. instances, independently of earlier rounds. The search rule supplies the next-token target $o_{\mathrm{cp}}(E)$ for each model-generated input prefix satisfying the conditions in Setting 3. Using the threshold ρ'' from that setting, we return the current model unchanged if $\|F(E)_{:,-1} - o_{\mathrm{cp}}(E)\|_2 < \rho''$ for every collected prefix. Otherwise, we add these input–target pairs to the accumulated dataset $\mathcal{D}_r$ and update the model by ERM of $\widehat{\mathcal{R}}_r(F) := \frac{1}{2|\mathcal{D}_r|}\sum_{(E,o)\in\mathcal{D}_r}\|F(E)_{:,-1} - o\|_2^2$. The sum includes all pairs collected across rounds. Inputs and targets remain fixed during each update.

In *teacher forcing* (Williams & Zipser, 1989), we instead use prefixes of N i.i.d. teacher traces as inputs. For prefix $E_{n,k}$ and next token $o_{n,k+1}$ from instance X_n, ERM minimizes $\widehat{\mathcal{R}}_{\mathrm{TF}}(F) := \frac{1}{2N}\sum_{n=1}^{N}\sum_{k<L(X_n)}\|F(E_{n,k})_{:,-1} - o_{n,k+1}\|_2^2$, where $L(X_n)$ counts next-token predictions on the teacher trace. We denote the corresponding population risk by $\mathcal{R}(F)$.

3 Backtracking is necessary

We compare single-path procedures with reward-ordered depth-first search to determine whether backtracking is needed for reliable search in our noisy binary-tree environment.[1]

Setting 1 (Search environment)**.** The environment $\mathcal{E}_H^{(C)}(\bar{\zeta})$ (Definition B.1) is a complete C-ary tree of depth $H \geq 2$ with ambiguity level $\bar{\zeta} \in (0, \frac{1}{2})$. Each internal node independently draws a uniform hidden preferred child $B_u \in [C]$, and the target leaf $s^\star$ is the leaf these preferences select. Rewards provide noisy evidence of these preferences: at a node u of depth $i-1$, $i \in \{1,\dots,H\}$, utilities $Y_{u,b} = \mathbf{1}\{b = B_u\} + \sigma_i \xi_{u,b}$ have independent standard Gaussian noise. We choose σ_i so that a non-preferred child outranks the preferred child with total probability exactly ζ_i. A fixed strictly increasing transform (Definition B.1) maps utilities to rewards in $(-1, 1)$, so $R = 1$; ties have probability zero. The misranking probabilities depend on height above the leaves: with $r_i = H - i + 1$ and $L_C = \lfloor \log_C H \rfloor$, $\zeta_i = \bar{\zeta}$ at the lowest L_C levels and $\zeta_i = H^{-3} C^{L_C - r_i}$ above them.

We count an *edge traversal* as a child selection or return to the parent. A *single-path forward procedure* moves only from parent to child and never revisits a node (Definition D.1). Procedures may observe the entire reward-labeled tree, but not the preferred children or target; acceptance is tested only at the committed node.

Theorem 1 (Separation between single-path and depth-first search)**.** *Under Setting 1 with $C = 2$, consider a random instance.*

- *(a)* Single-path success bound. *Let p be the posterior probability that slot 1 is preferred given the two rewards at a node with misranking probability $\bar{\zeta}$, and set $\kappa_2(\bar{\zeta}) := \mathbb{E}[p^2 + (1-p)^2] < 1 - \bar{\zeta}$. Define $C'_{\bar{\zeta}} := \kappa_2(\bar{\zeta})^{-1/2}$ and $\gamma_{\mathrm{obs}} := -\frac{1}{2}\log_2 \kappa_2(\bar{\zeta}) > 0$. Then every single-path forward procedure whose parameters and randomness are independent of the instance reaches $s^\star$ with probability at most $C'_{\bar{\zeta}} H^{-\gamma_{\mathrm{obs}}}$, even with the complete reward-labeled tree in view.*
- *(b)* Depth-first guarantee. *The reward-ordered depth-first walk reaches $s^\star$ within $5H$ edge traversals with probability at least $1 - H^{-3}$.*

Thus, regardless of architecture, training or computation budget, single-path success tends to zero as depth grows even with full reward observations, whereas backtracking succeeds with probability tending to one using a linear traversal budget. We next ask whether a Transformer can execute this search over its own continuous thoughts.

4 Existence of an exact search executor

We show that there exists a Transformer that executes backtracking search exactly over its own continuous thoughts.[2]

Setting 2 (Thought-interface execution)**.** Fix $C \geq 2$, $t \in (0, 1]$ and the stopped thought-interface problem with budget B_{sel}, so that $B = 2B_{\mathrm{sel}} + 1$ and the context has at most $T = (C+2)(B_{\mathrm{sel}} + 1)$

[1] Full statements: Theorem B.6; Lemma D.2.

[2] Full statement: Theorem E.8.

columns. Use raw feedback with thought clipping radius $Z \geq 1$, and Assumption 1. We call a prefix *canonical* if its discrete blocks have their prescribed values and inactive fields are zero; this condition does not constrain its thoughts. A canonical prefix is in scope if it satisfies the interface requirements (Definitions E.1 and E.2), is consistent with the search rule so far and, if a selection is due, has a reward gap at least t.

Theorem 2 (Exact search execution)**.** *Under Setting 2, we explicitly construct an executor with the following properties:*

(a) There is a causal Transformer F^{C} with one attention layer, $C + 2$ heads, width $\Theta_C(d + T)$, score radius $\log(T + 1)$ and FFN radius $O_{C,R,Z}(\sqrt{d + T} + t^{-1})$. It emits the search rule's next token exactly on every canonical prefix in scope, including both thoughts. Returns and termination require no reward-gap condition.

(b) Consequently, on every instance, its rollout from the root agrees with the rule until acceptance, the first selection with a reward gap below t (τ_t), termination, or budget exhaustion.

The construction realizes selection and backtracking as next-token predictions from the search history in context, rather than decisions made by an external search algorithm as in Tree of Thoughts (Yao et al., 2023). Attention retrieves the required information directly from this token sequence: C heads retrieve candidates, one retrieves the parent anchor, and a uniform head cancels the softmax denominator. The output FFN selects, returns or terminates, and exact copying of both thoughts prevents execution errors from accumulating before stopping. For fixed C, choosing $d = \Theta_C(H)$ and $B_{\mathrm{sel}} = O_C(H)$ gives width $\Theta_C(H)$, so the executor need not allocate one coordinate per tree node. The construction also handles learner-generated prefixes satisfying the conditions of Setting 3 (Theorem E.8(c)). These are existence guarantees; whether training produces an executor is the next question.

5 Learning to execute

We now ask when training yields correct search execution. We compare on-policy imitation learning on the learner's own histories with teacher forcing on fixed traces; for teacher forcing, we then compare raw feedback with full write-back.

Covering numbers. We use covering numbers to make the dependence of the learning guarantees on model complexity explicit. For each model class and prefix domain considered below, $\mathcal{N}(\alpha)$ is the minimum number of models from the class needed to approximate every member within α under $d_\infty(F, G) := \sup_E \|F(E)_{:,-1} - G(E)_{:,-1}\|_2$ (van der Vaart & Wellner, 1996). Lemma C.3 gives $\log \mathcal{N}(\alpha) \leq D \log(1 + A/\alpha)$ for $\alpha > 0$, with $D \geq 1$ proportional to the trainable parameter count and $A > 0$ depending on the architecture and the parameter and input norm bounds.

5.1 On-policy imitation learning

On-policy imitation learning supplies labels on the histories the model actually generates, so its targets follow the generated tree rather than a fixed teacher trajectory.[3]

Setting 3 (On-policy learning)**.** Under Setting 2, fix $N \geq 1$, $\delta_0 \in (0, 1)$ and a norm-bounded causal class $\mathcal{F}$ of fixed attention depth containing F^{C} at threshold t. Take $\mathcal{N}(\alpha)$ on all admitted prefixes below. The search rule specifies the next-token target $o_{\mathrm{cp}}(E)$ for the model's own prefix E. A prefix is *admitted* if its decoded history satisfies the interface requirements and is correct so far, its thoughts have norm at most Z, and no selection with a reward gap below t has occurred or is due. Each previous output must also have Euclidean error below $3(1 - \varepsilon_{\mathrm{emb}})/7$ in each of its current and parent embedding blocks, and below $1/4$ jointly in the remaining non-thought blocks, relative to the values prescribed by the search rule on its decoded input prefix. Set $\rho'' := \min\{1/4, 3(1 - \varepsilon_{\mathrm{emb}})/[14(C + 1)]\}$. Use the on-policy protocol of Section 2.2 with these targets and threshold ρ'', starting from an arbitrary fixed member of $\mathcal{F}$. Each round draws N new independent instances from the same fixed law. Check every admitted prediction before processing any stop or scope exit it triggers.

[3] Full protocol and proof: Definition E.11 and Theorem E.12.

Theorem 3 (On-policy learned execution)**.** *Under Setting 3, the protocol stops within* $K := \mathcal{N}(\rho''/3)$ *rounds and returns one Transformer* $\widehat{F}$*. With probability at least* $1 - \delta_0$ *over all training data, the probability of the following event on an independent test instance is at most*

$$1 \wedge \frac{\log \mathcal{N}(\rho''/3) + \log(1/\delta_0)}{N} :$$

$\widehat{F}$*, used at every step, violates closed-loop correctness at a step* $k \le B$ *before stopping, with* $k < \tau_t$*. The same bound covers leaving the admitted scope before the first of acceptance, termination, the* B*-step budget and a due selection with gap below* t*.*

Thus, on-policy imitation learning can learn search over continuous thoughts without quantization or write-back. Applying Lemma C.3 at $\alpha = \rho''/3$ gives $\log \mathcal{N}(\rho''/3) \le D \log(1 + 3A/\rho'')$, so the failure bound is D/N up to logarithmic factors. Here, N counts instances per round: total costs are at most NK instances and BNK labels. The bound $K \le (1 + 3A/\rho'')^D$ is exponential in D and does not guarantee efficient learning. We next ask whether teacher forcing on fixed traces can avoid repeated collection and labeling of learner rollouts and this round bound.

5.2 Teacher forcing

We study when teacher forcing on fixed traces yields correct search execution and how rounding predictions before storing them changes the guarantees.[4]

Quantized front end. Teacher forcing controls prediction errors on teacher traces, but these errors can change the candidates encountered during deployment. To stabilize candidate generation, we use a front end that rounds thoughts only for candidate lookup. This is analogous to treating synonymous expressions alike. On $\mathcal{E}_H^{(C)}(\bar{\zeta})$, let $\mathcal{A} = \{a_u\}$ be the node embeddings in $\mathbb{R}^d$. Define $Q(z) := \arg\min_{a_u \in \mathcal{A}} \|z - a_u\|_2$ with a fixed tie rule and $\rho := \frac{1}{2} \min_{u \ne v} \|a_u - a_v\|_2$. If $Q(z) = a_u$, the front end returns that node's child embeddings and observed rewards, with candidate thoughts equal to their child embeddings. Thus, its outputs are constant at distances less than ρ from each embedding, but the Transformer still reads the raw thoughts. The root has $z = z_{\mathrm{src}} = U = a_{\mathrm{root}}$.

5.2.1 Execution guarantees and limitations

Setting 4 (Teacher forcing)**.** Under Setting 1, use the quantized front end above, raw feedback and clipping radius $Z \ge 1$, with $B_{\mathrm{sel}} \ge 1$, $B = 2B_{\mathrm{sel}} + 1$ and $T = (C + 2)(B_{\mathrm{sel}} + 1)$; Assumption 1 holds. Before sampling at confidence $\delta \in (0, 1)$, fix a norm-bounded causal class $\mathcal{F}(\mathbf{\Lambda})$ of fixed attention depth, with architecture and norm bounds $\mathbf{\Lambda}$, in the original thought-interface layout (Definition E.1), with $C + 2$ heads. Take $\mathcal{N}(\alpha)$ on bounded numerical prefixes including the $\rho/2$ neighborhood of teacher prefixes, and assume an input Lipschitz bound $L_{\mathrm{in}} \ge 0$: $\|F(E)_{:,-1} - F(E')_{:,-1}\|_2 \le L_{\mathrm{in}} \|E - E'\|_{2,\infty}$, where the right-hand norm is the maximum column norm (appendix input-stability paragraph). Set $S_k(L_{\mathrm{in}}) := \sum_{j=0}^{k-1} L_{\mathrm{in}}^j$. Train on N independent traces $X_1, \ldots, X_N$ and define the *maximum training error* $\widehat{r}_{\max}(F) := \max_{i \le N} \max_{k < L(X_i)} \|F(E_k(X_i))_{:,-1} - o_{k+1}(X_i)\|_2$. The *trace margin* is the least eligible reward gap along a trace, with singleton or empty comparisons assigned $+\infty$. For the ERM guarantees, the class contains F^{C} at $t_N = \delta/[C(C-1)NB_{\mathrm{sel}}]$, with radii declared as functions of N, δ before sampling.

For Theorem 4(b), additionally require $B_{\mathrm{sel}} \ge H$, $a \in (0, \min\{1, \rho/2\}]$ and $Z \ge 1 + a$. Use the class of Definition A.4, containing F^{C} and closed under W_a (Definition F.6). This map introduces and amplifies a small difference between z_{src} and U during execution; its two internal coordinates are zero on inputs, targets and emissions.

Theorem 4 (Teacher forcing under front-end quantization)**.** *Under Setting 4:*

(a) Sufficient condition. *With probability at least* $1 - \delta$ *over the training sample, every member of* $\mathcal{F}(\mathbf{\Lambda})$ *with maximum training error at most* $\eta_* := \rho/[8S_B(L_{\mathrm{in}})]$ *reproduces the stopped teacher path with thoughts within* $\rho/2$*. On an independent test instance, this conclusion fails*

[4] Full statements: Theorems F.2 and F.8, Definition F.6, Lemma F.7, and Corollary F.9.

with probability at most

$$1 \wedge \frac{\log \mathcal{N}(\rho/[8S_B(L_{\mathrm{in}})]) + \log(1/\delta)}{N}.$$

Every ERM satisfies the same conclusion with $\log(2/\delta)$ *in place of* $\log(1/\delta)$.

(b) Counterexample. *For every* $\epsilon > 0$*, the counterexample class* $\mathcal{F}(\mathbf{\Lambda})$ *contains a member of teacher-forced risk at most*

$$\inf_{F \in \mathcal{F}(\mathbf{\Lambda})} \mathcal{R}(F) + \epsilon + \frac{(B+1)a^2}{4S_{H-1}(2)^2}.$$

On every instance whose stopped teacher trace reaches depth H*, its rollout leaves the teacher's decoded trace by the first selection of a depth-*H *node.*

Part (a) gives ERM execution guarantees from N fixed teacher traces, avoiding the rollout collection and exponential round bound of Theorem 3. However, errors bounded by r on teacher prefixes give rollout error at most $rS_k(L_{\mathrm{in}})$ after k steps. Applying Lemma C.3 at $\alpha = \rho/[8S_B(L_{\mathrm{in}})]$ gives $\log \mathcal{N}(\alpha) \le D \log(1 + 8AS_B(L_{\mathrm{in}})/\rho)$, hence a failure bound of $\widetilde{O}(DB/N)$ for fixed $L_{\mathrm{in}} > 1$. The sufficient error tolerance is then exponentially small in B. For fixed $L_{\mathrm{in}} < 1$, the failure bound is $\widetilde{O}(D/N)$ and the tolerance remains bounded below independently of B. Part (b) shows that, in the declared class, near-optimal teacher-forced risk can coexist with error accumulation and path failure. We next ask whether full write-back prevents this accumulation by rounding predictions before storage.

5.2.2 Full write-back

Keeping the same quantized front end, we use full write-back to prevent small output errors from accumulating across steps.[5] *Full write-back* stores decoded discrete blocks, clears unused candidate-embedding, reward and scratch blocks, and rounds z and z_{src} to their nearest node embeddings for selection and return outputs, or sets them to zero for termination outputs. Use Setting 4 with full write-back, $B_{\mathrm{sel}} = 5H$ and $\rho_{\mathrm{wb}} := \min\{\rho, 4/5\}$. The class is either the fixed sparse executor template with every supported coefficient trainable (Definition F.3, $D = O_C(H)$), or the dense companion with every coefficient of the same architecture trainable ($D = O_C(H^2)$). In both cases, choose radii before sampling so that the class contains F^{C} at t_N and $\log(1+A) = O_C(\log(NH/\delta))$ (Corollary F.4).

Theorem 5 (Robust execution with full write-back)**.** *Under Setting 4 with the full-write-back conditions above, fix* $0 \le \sigma \le \rho_{\mathrm{wb}}/4$ *before sampling. With probability at least* $1 - \delta$ *over the training sample, every ERM, and more generally every model* F *in the class satisfying* $\widehat{r}_{\max}(F) + \sigma \le \rho_{\mathrm{wb}}/4$*, has a full-write-back rollout equal to the teacher's stopped trace, with every stored column equal to its teacher counterpart, even when each predicted token is perturbed before decoding by a history-dependent vector of Euclidean norm at most* σ*. The front end, supplied metadata and stored context remain unperturbed. On an independent test instance, this conclusion fails with probability at most*

$$v_N := 1 \wedge \frac{\log \mathcal{N}(\rho_{\mathrm{wb}}/8) + \log(2/\delta)}{N}.$$

Hence its decoded path reaches $s^\star$ *within* $5H$ *child selections with probability at least* $1 - H^{-3} - v_N$*, where* $v_N = \widetilde{O}_C(H/N)$ *for the sparse template and* $v_N = \widetilde{O}_C(H^2/N)$ *for the dense companion.*

Remark (Relation to the counterexample). Full write-back also provides a path-success guarantee for the counterexample model in Theorem 4(b). For fixed a and unchanged model parameters, its path-failure probability is bounded by $O_C(\delta/N + \epsilon + H4^{-H})$, even under the output perturbations specified in Theorem 5 with $\sigma \le \rho_{\mathrm{wb}}/4$ (Corollary F.10).

Full write-back removes $S_B(L_{\mathrm{in}})$ from the covering scale and gives horizon-independent tolerance for prediction errors and perturbations by restoring every stored teacher token. Applying Lemma C.3 at $\alpha = \rho_{\mathrm{wb}}/8$ gives $\log \mathcal{N}(\alpha) \le D \log(1 + 8A/\rho_{\mathrm{wb}})$, so the execution-failure bound v_N is D/N up to logarithmic factors. Substituting the sparse and dense dimensions $D = O_C(H)$ and $D =$

[5] Full statement and proof: Theorem F.5.

$O_C(H^2)$ yields the stated rates, compared with raw-feedback upper bounds of $\widetilde{O}_C(H^2/N)$ and $\widetilde{O}_C(H^3/N)$ under the same norm bounds (Theorem 4(a)). However, raw feedback can retain distinct continuous thoughts within the same front-end lookup region, whereas full write-back maps them to the same node embedding. These stronger guarantees thus come at the cost of restricting stored thoughts to a finite set.

6 Experiments

To complement the ERM guarantees, we test whether standard optimization of teacher-forced loss yields successful search. We also evaluate the same trained models under raw feedback and full write-back to assess robustness to the prediction errors remaining after training.

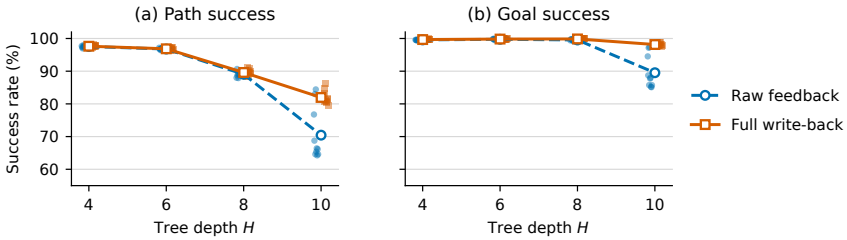


Figure 2: Path and goal success rates under raw feedback and full write-back across tree depths. Lines show ten-seed means; smaller markers show individual seeds with horizontal offsets.

We use the binary environment of Setting 1 with $\bar{\zeta} = 0.2$ and the quantized front end of Section 5.2. At each $H \in \{4, 6, 8, 10\}$, we train ten randomly initialized one-attention-layer Transformers on 4096 teacher traces using $20{,}000$ AdamW updates. At each depth, we evaluate all ten models under both protocols on the same 2048 test instances.

Figure 2 shows high goal success after AdamW training, with full write-back achieving mean goal success above 98% at each depth in this grid. The two feedback protocols perform similarly at $H \leq 8$, but differ more substantially at $H = 10$. At that depth, full write-back raises mean path success from 70.43% to 81.97% and goal success from 89.53% to 98.15%. At $H = 10$, path success improves for all ten seeds. Per-seed outcomes and paired uncertainty appear in Appendix G.2.

These results suggest that standard optimization can learn successful search in these environments, while the gains from full write-back support the role of feedback storage in robustness to remaining prediction errors.

7 Conclusion and Limitations

Conclusion. Backtracking enables reliable search in our noisy binary-tree family (Theorem 1), and a one-attention-layer Transformer can execute it autoregressively over its own continuous thoughts (Theorem 2). Such execution can be learned through on-policy imitation learning without quantizing thoughts (Theorem 3), but data collection costs motivate teacher forcing on fixed traces. With a quantized front end, teacher forcing also gives ERM execution guarantees (Theorem 4(a)). However, raw feedback can amplify errors, allowing near-optimal teacher-forced risk to coexist with path failure in the declared counterexample class (Theorem 4(b)). Full write-back prevents this accumulation and gives horizon-independent tolerance for prediction errors and output perturbations (Theorem 5), at the cost of restricting stored thoughts to a finite set. Together, these results establish autoregressive backtracking search execution and learning by Transformers under the stated conditions, with guarantees shaped by supervision and feedback storage. Experiments suggest successful learning through standard optimization and improved robustness with full write-back.

Limitations. We study search control with a fixed proposer, leaving joint learning of candidate generation and search open. The exact-execution and on-policy guarantees apply only before a selection with a reward gap below the specified threshold. Our on-policy round bound is exponential in the trainable parameter count. Our ERM guarantees do not cover gradient-based optimization, finite-precision computation or layer-normalized Transformers.

REPRODUCIBILITY STATEMENT

All theoretical claims are proved in the appendices: Appendices A–F carry the full statements and proofs, and Appendix H is a correspondence table locating each theoretical main-text claim among the full statements. Experiments: Appendix G specifies the environment, thought-interface conventions, model, training and validation selection, the two feedback protocols, paired evaluation, and all forty per-seed results. The result tables are generated from the evaluation records by `scripts/summarize_e1_paper.py`; `scripts/plot_e1_paper.py` generates the main figure from the same summary. The anonymous supplementary code reproduces this experiment and generates the result tables and figure, with dependencies and commands in its README.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript drafting and editing, literature discovery, research ideation, experimental design and implementation, interpretation of results, and developing proof strategies and writing proofs. The authors directed the research and made the substantive decisions on its formulation, methods, and claims. We evaluated AI-suggested references against the original publications and selected the works cited. For the theoretical results, we supplied and refined proof ideas, checked the mathematical arguments against the stated assumptions, and corrected deficiencies identified during review. All AI-assisted material included in the paper was critically evaluated and approved by the authors. We take full responsibility for the final manuscript, theoretical claims, code, and experimental results.

REFERENCES

Riccardo Alberghi, Elizaveta Demyanenko, Luca Biggio, and Luca Saglietti. On the bias of next-token predictors toward systematically inefficient reasoning: A shortest-path case study. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5910–5937, 2025.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 2296–2318, 2024.

Artur Back de Luca and Kimon Fountoulakis. Simulation of graph algorithms with looped transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2319–2363, 2024.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 7654–7664, 2022.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

Kanishk Gandhi, Denise H J Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah Goodman. Stream of search (SoS): Learning to search in language. In *First Conference on Language Modeling*, 2024.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017. URL `https://arxiv.org/abs/1704.00805v2`.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E. Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025.

Joongwon Kim, Anirudh Goyal, Liang Tan, Hannaneh Hajishirzi, Srinivasan Iyer, and Tianlu Wang. ASTRO: Teaching language models to reason by reflecting and backtracking in-context. *arXiv preprint arXiv:2507.00417*, 2025.

Jungtaek Kim, Thomas Zeng, Ziqian Lin, Minjae Lee, Chungpa Lee, Jy-yong Sohn, Hyung Il Koo, and Kangwook Lee. Transformers in the dark: Navigating unknown search spaces via bandit feedback. *Transactions on Machine Learning Research*, 2026.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lucas Lehnert, Sainbayar Sukhbaatar, DiJia Su, Qinqing Zheng, Paul McVay, Michael Rabbat, and Yuandong Tian. Beyond A*: Better planning with transformers via search dynamics bootstrapping. In *First Conference on Language Modeling*, 2024.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Harsh Patel. A verifiable search is not a learnable chain-of-thought. *arXiv preprint arXiv:2606.21884*, 2026.

Yin Jun Phua, Tony Ribeiro, Tuan Nguyen, and Katsumi Inoue. Can transformers learn to verify during backtracking search? *arXiv preprint arXiv:2605.22221*, 2026.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Abulhair Saparov, Srushti Ajay Pawar, Shreyas Pimpalgaonkar, Nitish Joshi, Richard Yuanzhe Pang, Vishakh Padmakumar, Seyed Mehran Kazemi, Najoung Kim, and He He. Transformers struggle to learn to search. In *The Thirteenth International Conference on Learning Representations*, 2025.

Max Simchowitz, Daniel Pfrommer, and Ali Jadbabaie. The pitfalls of imitation learning when actions are continuous. In *Proceedings of the Thirty Eighth Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5248–5351, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996. doi: 10.1007/978-1-4757-2545-2.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed 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.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Agentic transformers provably learn to search via reinforcement learning. *arXiv preprint arXiv:2606.00183*, 2026a.

Xiao-Wen Yang, Xuan-Yi Zhu, Ding-Chu Zhang, Wen-Da Wei, Jie-Jing Shao, Zhi Zhou, Lan-Zhe Guo, and Yu-Feng Li. Step back to leap forward: Self-backtracking for symbolic reasoning and planning in language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (33):27657–27665, 2026b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 62138–62160, 2024.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart J. Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Emergence of superposition: Unveiling the training dynamics of chain of continuous thought. In *The Fourteenth International Conference on Learning Representations*, 2026.

GUIDE TO THE APPENDICES

The appendix orders proofs by technique rather than by theorem number; Table 5 maps each main theorem to its sources. Appendix A defines dictionaries, Transformer classes, teacher-forced risks and building blocks; Definition A.2 specifies the causal class. Appendix B proves C-ary noise and reward bounds and the $5H$ DFS bound, with the binary posterior-moment upper bound as its $C = 2$ example. Appendix C gives the covering-number bounds and decoding under bounded errors. Appendix D assembles Theorem 1, defines single-path forward procedures and shows that the single-path bound also permits arbitrary interleaved computation. Appendix E defines the thought interface and proves exact execution and on-policy learning (Theorems 2 and 3); the executor theorem separates canonical prefixes, rollout closure and admitted raw prefixes, after defining copy labels by finite replay. Definition E.11 and Theorem E.12 prove Theorem 3 for a shared Transformer, including finite termination and ERM selection. Appendix F proves the sufficient condition, including the ERM clause, and original-interface counterexample of Theorem 4(a) and (b), respectively, under Setting 4. Its Subsection F.2 defines the fixed sparse executor template and proves its $O_C(H)$ trainable parameter count and covering-number bound in Corollary F.4. The full-write-back clause of Theorem F.2, including its ERM sentence, Theorem F.5 and Lemma B.8 establish the constant-tolerance execution and goal guarantees of Theorem 5 under Setting 4 with the full-write-back conditions of Section 5.2.2. Lemma F.1 proves restoration under bounded, history-dependent output perturbations; Corollary F.10 applies it to the same counterexample model as in Theorem 4(b). Appendix G gives the standard-training protocol, the paired raw/full-write-back comparison, all per-seed results and reproduction commands. Appendix H locates the five main theorems' proofs.

A TOKEN INTERFACES, TRANSFORMER CLASSES, AND BUILDING BLOCKS

This appendix defines sign dictionaries, Transformer classes, teacher-forced risks and ReLU operations. Appendix D defines single-path forward procedures and proves the single-path success bound; Appendix E.1 constructs the thought executor. Appendices C and F give the covering-number bounds and learning guarantees.

A.1 SIGN DICTIONARIES

Throughout, $\sigma(t) := \max\{t, 0\}$ denotes the ReLU, applied coordinatewise, and the incoherence of the node dictionary is written ε inside the appendices (it is the $\varepsilon_{\mathrm{emb}}$ of the main text; every other use of ε carries a subscript).

Lemma A.1 (Rademacher sign dictionaries). *For every $\varepsilon \in (0, 1/2)$ and every finite symbol set $\mathcal{Q}$ with $|\mathcal{Q}| \le S+2$ there exist sign embeddings $a_q \in \{\pm 1/\sqrt{d}\}^d$ with $\langle a_q, a_q \rangle = 1$ and $|\langle a_q, a_{q'} \rangle| \le \varepsilon$ for $q \neq q'$, at dimension*

$$d = \left\lceil \frac{2}{\varepsilon^2} \log\big(2(S+2)(S+1)\big) \right\rceil = O(\varepsilon^{-2} \log S);$$

for trees of branching factor at most $C \ge 2$ and depth at most H, $d = O(\varepsilon^{-2} H \log C)$ suffices. For the stopped thought problem, positions and positional content pointers use the orthonormal dictionary $e_1, \ldots, e_T$ with $T = (C+2)(B_{\mathrm{sel}}+1)$; $d_{\mathrm{pos}} = T = O_C(H)$ at $B_{\mathrm{sel}} = 5H$. The node dictionary remains the sign dictionary.

Proof. Let $n := |\mathcal{Q}| \le S+2$. For each symbol $q \in \mathcal{Q}$ and each coordinate $\ell \in \{1, \ldots, d\}$, draw independent Rademacher signs $\sigma_{q,\ell} \in \{-1, +1\}$ (each value with probability $\frac{1}{2}$) and set $a_q := d^{-1/2}(\sigma_{q,1}, \ldots, \sigma_{q,d})^\top$. Then $a_q \in \{\pm 1/\sqrt{d}\}^d$ and $\langle a_q, a_q \rangle = d^{-1} \sum_\ell \sigma_{q,\ell}^2 = 1$ deterministically. For $q \neq q'$, the products $X_\ell := \sigma_{q,\ell}\sigma_{q',\ell}$, $1 \le \ell \le d$, are independent Rademacher variables and $\langle a_q, a_{q'} \rangle = d^{-1} \sum_\ell X_\ell$, so Hoeffding's inequality (Vershynin, 2018, Section 2.2) gives

$$\mathbb{P}\big(|\langle a_q, a_{q'} \rangle| > \varepsilon\big) = \mathbb{P}\Big(\Big|\sum_{\ell=1}^{d} X_\ell\Big| > d\varepsilon\Big) \le 2e^{-d\varepsilon^2/2}.$$

A union bound over the $n(n-1)/2 \le (S+2)(S+1)/2$ unordered pairs bounds the probability that some pair violates the incoherence bound by $(S+2)(S+1)e^{-d\varepsilon^2/2}$, which is at most $\frac{1}{2}$ for the displayed d. Hence some realization of the signs satisfies every requirement, and $d \le 2\varepsilon^{-2}\big(2\log(S+2) + \log 2\big) + 1 = O(\varepsilon^{-2} \log S)$. For a tree of branching factor at most C and depth at most H, the number of nodes is at most $1 + C + \cdots + C^H \le 2C^H$ (for $C \ge 2$), so $\log(S+2) = O(H \log C)$ and the second bound follows from the first. The stated positional dimension follows directly from the size of the orthonormal dictionary. □

Dependence on incoherence and clipping radius. Fix $0 < \varepsilon = \varepsilon_{\mathrm{emb}} < 1/2$ and a finite clipping radius Z. Asymptotic constants may depend on the fixed $C, \varepsilon_{\mathrm{emb}}, Z$, and on R when rewards are not normalized. The formulas used below are

$$\sqrt{(1-\varepsilon)/2} \le \rho \le \sqrt{(1+\varepsilon)/2}, \quad \eta = \frac{3(1-\varepsilon)}{7}, \quad \rho'' = \min\left\{\frac{1}{4}, \frac{\eta}{2(C+1)}\right\}.$$

Here, $\rho > 1/2$, $\rho/2 < 1/2$, and $\eta < 1/2 < \rho$. The threshold ramp's transition interval has width $(1-\varepsilon)/7 > 1/14$ and slope $7/(1-\varepsilon) < 14$. For the full-write-back guarantee, set

$$\rho_{\mathrm{wb}} = \min\{\rho, 4/5\}.$$

Full write-back uses $5\rho_{\mathrm{wb}}/8 \le 1/2$, with strict prediction-error control at equality, as proved in Theorem F.2. The quantized front end needs $Z \ge 1$; the modification W_a uses $Z \ge 1 + a$. All sufficient radii below include these fixed-Z costs.

A.2 Token interfaces and teacher prefixes

A finite tree $\mathcal{T}$ with $|\mathcal{V}(\mathcal{T})| \le S$ carries directed parent-to-child edges $\mathcal{E}(\mathcal{T})$, a root with a dummy parent 0, and an observed reward $r_{\mathcal{T}}(v)$ with $|r_{\mathcal{T}}(v)| \le R$ on every node. In the environment of Definition B.1, the rewards are the transformed rewards of Lemma B.3 with $R = 1$ and sibling gaps are almost surely positive. Exact execution statements specify a positive selection threshold. Each discrete symbol is embedded by the sign dictionary of Lemma A.1.

Problem TI (DFS over an implicit generated tree; TI abbreviates thought interface). The search history and its links are stored in the context tokens, with no separate search memory. The token dimension is p_{TI}. Write $o_{\mathrm{TI},0}, \ldots, o_{\mathrm{TI},L_{\mathrm{TI}}(\mathcal{T})}$ for the root and teacher control tokens. At prediction step k, $E_{\mathrm{TI},k}(\mathcal{T})$ contains the root, the first k control outputs and all intervening interface-written records; its target is $o_{\mathrm{TI},k+1}$. Teacher forcing uses these teacher tokens rather than model

predictions. Let $L_{\max,\mathrm{TI}}$ bound the number of prediction steps on the training support. A fixed columnwise front end

$$g_{\mathrm{pre}} : (z, a_u, e_b) \longmapsto \big(z_b(z,u),\, a_{v_b(z,u)},\, r_b(z,u)\big)$$

generates, from the thought z attached to the current node u and a slot indicator e_b with $b \in [C]$, the candidate thought, child embedding, and reward. Invalid proposals are $(0_d, a_{\mathrm{eos}}, 0)$; valid thoughts satisfy $\|z\|_2 \le Z$, rewards $|r| \le R$, and the valid proposals may occupy any subset of the C slots, recorded by validity bits and their total count. The token layout, search rule, and stopped problem with budget B_{sel} are fixed in Appendix E.1 (Definitions E.1, E.2, E.3); the interface writes the C candidate records $g_{\mathrm{pre}}(z, \mathrm{Dec}_U(U), e_b)$ after every fresh selected-state token before stopping. Under the fresh-embedding condition of Definition E.2, each of at most S nodes is entered once and receives at most C candidate records. Each of the at most $S-1$ selected edges is returned along at most once. Including a possible terminal output gives $L_{\max,\mathrm{TI}} \le 2S-1$ prediction steps and at most $(C+2)S$ columns. Without that condition, a repeated embedding opens a new occurrence; the budget bounds the prediction steps by $2B_{\mathrm{sel}}+1$ and the columns by $(C+2)(B_{\mathrm{sel}}+1)$. The executor is constructed in Appendix E.

A.3 Transformer classes and teacher-forced risks

For a token matrix $X = [x_1, \dots, x_m]$, a J-head residual attention layer is

$$\mathcal{A}_{\mathbf{V},\mathbf{B}}(X) := X + \sum_{h=1}^{J} V_h X \,\mathrm{softmax}(X^\top B_h X),$$

softmax columnwise, B_h the combined key–query score matrix, with the norm-bounded attention class $\mathfrak{A}_p(J, \Lambda_V, \Lambda_B) := \{\mathcal{A}_{\mathbf{V},\mathbf{B}} : \|V_h\|_{\mathrm{op}} \le \Lambda_V, \|B_h\|_{\mathrm{op}} \le \Lambda_B\}$. A ReLU FFN class $\Psi_p(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$ consists of depth-L_{ffn}, width-W_{ffn} maps $\psi : \mathbb{R}^p \to \mathbb{R}^p$ with per-layer weight and bias norms bounded by Λ_{ffn} and the linear growth bound $\|\psi(x)\|_2 \le c_{\mathrm{ffn}}\|x\|_2$ (the precise parametrization is fixed at the start of Appendix C); FFN layers act columnwise by replacing their input (no FFN skip connection). The architecture and norm bounds are $\mathbf{\Lambda} = (L_{\mathrm{att}}, J, L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_V, \Lambda_B, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$, and the Transformer class is the alternation

$$\mathcal{F}_{\mathrm{TI}}(\mathbf{\Lambda}) := \Big\{ \mathcal{F}_{\psi_{L_{\mathrm{att}}}} \circ \mathcal{A}_{L_{\mathrm{att}}} \circ \cdots \circ \mathcal{F}_{\psi_1} \circ \mathcal{A}_1 \circ \mathcal{F}_{\psi_0} \;\Big|\; \mathcal{A}_r \in \mathfrak{A}_{p_{\mathrm{TI}}}(J, \Lambda_V, \Lambda_B),\ \psi_r \in \Psi_{p_{\mathrm{TI}}}(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}}) \Big\}$$

with the fixed front end g_{pre} applied by the interface (not trainable, excluded from the model's parameter and norm bounds). All attention scores are bilinear, $q(y,x) = y^\top B x$: every score term that reads only the source y is multiplied by a query-side gate coordinate (the constant register $x_\iota = 1$, or a state flag that equals one on the queries for which the head is active), so all displayed scores are realizable as $y^\top B x$. Two remarks. First, the base class contains *no layer normalization*; this is a strong architectural simplification (Chiang & Cholak, 2022), and every statement below is for the LN-free classes. Second, the learning guarantees concern ERM over these norm-bounded classes. The execution guarantees also apply to any class member satisfying their stated error bounds, regardless of how it was obtained; they do not show that SGD attains those bounds or an ERM. The causal class $\mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}$ (Definition A.2) is obtained by causal masking and concatenated positional blocks and is the model used by the execution and learning results.

The teacher-forced population risk of F is

$$\mathcal{R}_{\mathrm{TI}}(F) := \tfrac{1}{2}\,\mathbb{E}_{\mathcal{T}} \sum_{k=0}^{L_{\mathrm{TI}}(\mathcal{T})-1} \big\| F(E_{\mathrm{TI},k}(\mathcal{T}))_{:,-1} - o_{\mathrm{TI},k+1}(\mathcal{T}) \big\|_2^2,$$

with empirical version $\widehat{\mathcal{R}}_{\mathrm{TI},N}$ on N i.i.d. instances and ERM $\widehat{F}_{\mathrm{TI},N}$. The teacher's next-token target is a deterministic function of its prefix, with fixed tie rules; the search environment is almost surely tie-free. Thus the unrestricted Bayes risk is zero, but we do not assume that the infimum over a fixed norm-bounded Transformer class is zero. In the causal class, the last column's own positional block is supplied with the input and is not predicted: throughout, $F(E_{\mathrm{TI},k})_{:,-1}$ and $o_{\mathrm{TI},k+1}$ are compared after the fixed coordinate projection that deletes that block, a nonexpansive map that changes no bound below; backlink blocks are predicted content, decoded for execution and written back only under a write-back protocol.

A.4 CAUSAL TRANSFORMERS

Definition A.2 (Causal decoder Transformer)**.** The causal class $\mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}(\mathbf{\Lambda})$ is obtained from $\mathcal{F}_{\mathrm{TI}}(\mathbf{\Lambda})$ by replacing every attention layer with its causally masked version: column i attends only to columns $i' \le i$,

$$\mathcal{A}^{\mathrm{c}}_{\mathbf{V},\mathbf{B}}(X)_{:,i} := x_i + \sum_{h=1}^{J} V_h X_{:,\le i}\,\mathrm{softmax}\big(X^{\top}_{:,\le i} B_h\, x_i\big).$$

Every input column carries a concatenated positional block. For the stopped thought problem on at most $T = (C+2)(B_{\mathrm{sel}}+1)$ columns, $\pi_i = e_i \in \mathbb{R}^T$, and its backlink and fresh-identifier content blocks use this same dictionary with zero for null. Own-position blocks are supplied input metadata and excluded from prediction loss; positional content blocks remain predicted. Positional dictionaries are fixed before sampling.

Lemma A.3 (Prefix consistency: recomputation equals KV-caching)**.** *For every $F \in \mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}$ and every prefix $X = [x_1, \dots, x_n]$, the representation of column i after any number of layers is a function of $x_{\le i}$ only. Consequently, appending columns never changes earlier representations, and per-step full recomputation coincides with cached autoregressive decoding.*

Proof. Induction over layers: the input layer satisfies the claim; a causally masked attention output at column i depends on the layer inputs at columns $\le i$; FFN layers act columnwise. Composition preserves the property, and the caching statement is its restatement. □

Definition A.4 (The class used for the counterexample)**.** For Setting 4, fix a as there, set $u := a/S_{H-1}(2)$, and fix a causal base class $\mathcal{K}(\mathbf{\Lambda}_0)$ on the original p coordinates, with architecture and norm bounds at least those of the exact executor in Theorem 2 at the declared threshold. Its bounds are fixed before sampling; any fixed sparse support includes the executor's support. Set $p' = p + 2$. Append two reserved coordinates, with zero inputs and targets. Every base FFN acts as $(x, r_1, r_2) \mapsto (\psi(x), r_1, r_2)$: it neither reads nor changes the reserved coordinates. All attention score and value matrices have zero rows and columns there. Thus attention skip connections preserve them. Only the additional input and output FFN stages may write them; every emitted column has both coordinates reset to zero.

A Transformer has a declared representation $G = F_{G_0,b}$, with $G_0 \in \mathcal{K}(\mathbf{\Lambda}_0)$ and $b \in \{0,1\}$ a fixed architecture choice, not an input or predicted bit. There are fixed slots before the base Transformer's initial FFN and after the base output FFN. For $b = 0$, they are the identity on original coordinates. For $b = 1$, the first slot writes $q = (z_{\mathrm{src}} - U)_1$ and $\kappa = s_{\mathrm{bit}}(\sum_i \pi_i)$ into the reserved coordinates before any base operation, and the last slot applies Definition F.6 to G_0's output using q, κ. Here, s_{bit} is Lemma A.6(v). Declare

$$\mathcal{F}(\mathbf{\Lambda}) := \{F_{G_0,b} : G_0 \in \mathcal{K}(\mathbf{\Lambda}_0),\ b \in \{0,1\}\}, \qquad W_a(F_{G_0,b}) := F_{G_0,1}.$$

Applying W_a replaces these fixed slots; it does not append another copy of the output operation to an already modified Transformer. This specification on representations is part of the class definition.

The declared widths, depths and radii exceed those needed by F^{C} and the base class by the additional FFN costs. For explicit sufficient allowances, let L_0, W_0, Λ_0, c_0 be the base FFN bounds and B_0 the base class's output envelope on the predeclared input ball of radius B_x from Lemma C.1. Set $M_{\mathrm{core}} = \max\{2, B_0\}$ and take

$$\begin{aligned} L_{\mathrm{ffn}} &= L_0 + 32, \qquad W_{\mathrm{ffn}} = W_0 + 64p', \\ \Lambda_{\mathrm{ffn}} &\ge \max\{2\Lambda_0 + 2, K_C((1 + M_{\mathrm{core}})\sqrt{p'} + a^{-1})\}, \\ c_{\mathrm{ffn}} &\ge K_C(1 + c_0)(1 + a^{-1})\sqrt{p'}. \end{aligned}$$

Here, K_C is a fixed sufficiently large constant; the fixed sparse support adds at most $K_C(L_{\mathrm{att}} + 1)(L_0 + 32)p'$ nonzeros. Attention depth, head count and score and value radii stay those of the base class. Unused stages are padded by signed ReLU identities. The slots and their support cannot be used by base parameters. Lemma F.7 proves membership, closure, the loss identity and the inherited cover and input bounds for this subclass.

Remark A.5 (Two attention conventions, and terminology). Two conventions for attention appear. The causal class $\mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}$ re-encodes nothing: column i attends to columns $i' \le i$, and per-step recomputation equals KV-cached autoregressive decoding (Lemma A.3); it is the model of all learned

executors, and may be described as a causally masked decoder-only Transformer with concatenated absolute positional encodings; the thought executor of Appendix E and the counterexample of Theorem F.8 belong to it as well. The class $\mathcal{F}_{\mathrm{TI}}$ re-encodes the entire prefix at every step; it is used only as the ambient class of the cover lemma and is never needed by a construction. Neither class is a standard residual pre-LN Transformer: FFNs are direct overwrites and the base class has no layer normalization; the interface is autoregressive (the prediction at step k is a function of $E_{\mathrm{TI},k}$ only) and the model's own emitted tokens are the search memory.

A.5 THE FFN GADGET LIBRARY

Every FFN used in this paper is assembled from the finitely many columnwise maps of the next lemma. Each is written as an explicit ReLU network of the class $\Psi_p(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$ of Appendix A.3, i.e. as $\psi = (A_L\sigma(\cdot) + b_L) \circ \cdots \circ (A_1 \cdot + b_1)$ with all hidden widths at most W_{ffn}, $\max_\ell\{\|A_\ell\|_{\mathrm{op}}, \|b_\ell\|_2\} \le \Lambda_{\mathrm{ffn}}$, and $\|\psi(x)\|_2 \le c_{\mathrm{ffn}}\|x\|_2$ for every x (so in particular $\psi(0) = 0$). A binary register is a coordinate whose value on every legal token lies in $\{0, 1\}$; a sign embedding is a block in $\{\pm 1/\sqrt{d}\}^d$; "exact on D" means that the displayed ReLU expression equals the target map on every input in D.

Lemma A.6 (FFN gadget library)**.** *Let $M_0 > 0$. The following maps are exact on the stated domains, and each is realized by a single hidden ReLU layer whose input coefficients are affine in the input coordinates.*

(i) Copy under a binary gate. For $\chi \in \{0, 1\}$ and v with $\|v\|_\infty \le M_0$, coordinatewise,

$$\chi \cdot v = \sigma\big(v - M_0(1 - \chi)\big) - \sigma\big(-v - M_0(1 - \chi)\big).$$

(ii) Test a threshold on margin-separated inputs. For $\theta_0 < \theta_1$,

$$f_{\theta_0,\theta_1}(x) := \sigma\Big(\frac{x - \theta_0}{\theta_1 - \theta_0}\Big) - \sigma\Big(\frac{x - \theta_1}{\theta_1 - \theta_0}\Big)$$

equals 0 for $x \le \theta_0$ and 1 for $x \ge \theta_1$, is nondecreasing and $(\theta_1 - \theta_0)^{-1}$-Lipschitz on $[\theta_0, \theta_1]$, and satisfies $f(0) = 0$ whenever $\theta_0 \ge 0$ (the identity holds for every $\theta_0 < \theta_1$, but membership in Ψ_p requires $\theta_0 \ge 0$; every use in this paper satisfies it). It realizes the indicator $\mathbf{1}\{x \ge \theta\}$ exactly on inputs with $x \le \theta - \mu$ or $x \ge \theta$, by taking $(\theta_0, \theta_1) = (\theta - \mu, \theta)$; no claim is made on the band (θ_0, θ_1).

(iii) Shift one-hot vectors and combine bits. For a one-hot block $h \ge 0$ and a fixed $0/1$ matrix Π (a permutation or a shift), $\Pi h = \sigma(\Pi h)$. For binary registers $b_1, \dots, b_k$,

$$b_1 \wedge \cdots \wedge b_k = \sigma\Big(\sum_i b_i - (k - 1)\Big), \qquad b_1 \vee b_2 = \sigma(b_1 + b_2) - \sigma(b_1 + b_2 - 1),$$

exactly on $\{0, 1\}^k$. A gate that is a conjunction of binary registers is folded into the bias of (i): with $\chi := b_1 \wedge \cdots \wedge b_k$, the expression $\sigma\big(v - M_0(k - \sum_i b_i)\big) - \sigma\big(-v - M_0(k - \sum_i b_i)\big)$ equals $\chi \cdot v$ on $\{0, 1\}^k \times [-M_0, M_0]$, so no second layer is needed. Complements $1 - b$ are never emitted; they enter only as bias shifts $M_0(1 - b)$.

(iv) Round a sign embedding. With

$$q_{\mathrm{sgn}}(t) := -\frac{1}{\sqrt{d}} + 2\sigma\Big(t + \frac{1}{2\sqrt{d}}\Big) - 2\sigma\Big(t - \frac{1}{2\sqrt{d}}\Big),$$

one has $q_{\mathrm{sgn}}(t) = +1/\sqrt{d}$ for $t \ge 1/(2\sqrt{d})$, $q_{\mathrm{sgn}}(t) = -1/\sqrt{d}$ for $t \le -1/(2\sqrt{d})$, $q_{\mathrm{sgn}}(0) = 0$, and $|q_{\mathrm{sgn}}(t)| \le 2|t|$; hence the coordinatewise map q_{sgn} sends every vector within ℓ_∞-distance $1/(2\sqrt{d})$ of a sign embedding $a \in \{\pm 1/\sqrt{d}\}^d$ exactly to a (the constant $-1/\sqrt{d}$ is the output bias b_2, and $q_{\mathrm{sgn}}(0) = 0$).

(v) Round a bit. $b_{\mathrm{rnd}}(t) := \sigma(2t - \frac{1}{2}) - \sigma(2t - \frac{3}{2}) = f_{1/4,3/4}(t)$ equals 0 for $t \le \frac{1}{4}$ and 1 for $t \ge \frac{3}{4}$: every value within $\frac{1}{4}$ of a bit is rounded to that bit.

(vi) Write a constant under a binary gate. For a fixed vector c with $\|c\|_\infty \le M_0$ and a binary register χ, $\chi \cdot c = \sigma\big(c - M_0(1 - \chi)\big) - \sigma\big(-c - M_0(1 - \chi)\big)$ coordinatewise ((i) applied to the constant c; the ReLU inputs are affine in χ alone).

(vii) Restore a blank or legal token. For a binary flag χ, a block v with $\|v\|_\infty \le M_0$, and a fixed canonical token c with $\|c\|_\infty \le M_0$,

$$(1-\chi)\cdot v+\chi\cdot c = \Big[\sigma(v-M_0\chi)-\sigma(-v-M_0\chi)\Big]+\Big[\sigma\big(c-M_0(1-\chi)\big)-\sigma\big(-c-M_0(1-\chi)\big)\Big],$$

which passes v through when $\chi = 0$ and overwrites it with c when $\chi = 1$.

Resource bounds for compositions. *Fix $C \ge 2$, $R, Z > 0$, and $t \in (0, 1]$. A sequence of $O_C(1)$ stages retaining $O_C(p)$ token coordinates for later use has constant affine depth, hidden width $O_C(p)$, and $O_C(p)$ nonzero weights and biases per affine layer if each stage consists of coordinatewise gadgets, bounded-arity Boolean operations, partial coordinate permutations, or $O_C(1)$ linear reductions or broadcasts. Keep reductions, scalar ramps, and broadcasts in separate affine layers. If all coefficients are $O_{C,R,Z}(1)$ except for $O_C(1)$ scalar ramp coefficients of size $O_C(t^{-1})$, then*

$$\max_i\{\|W_i\|_{\mathrm{F}}, \|b_i\|_2\} = O_{C,R,Z}(\sqrt{p} + t^{-1}), \qquad \|W_i\|_{\mathrm{op}} \le \|W_i\|_{\mathrm{F}}.$$

For linear growth, use $\Gamma_A(g, v) := \sigma(v - A(1-g)) - \sigma(-v - A(1-g))$, clamp scalar control gates to $[0, 1]$, and guard each constant block c of bounded norm by $\kappa = \mathrm{clip}_{[0,1]}(\iota)$ or $\kappa = s_{\mathrm{bit}}(\iota)$, with $s_{\mathrm{bit}} = b_{\mathrm{rnd}}$ of (v), coordinatewise with magnitude $|c_i|$ (zero coordinates are omitted). Suppose there are $O_C(1)$ other emitted blocks, each obtained from raw blocks by $O_C(1)$ copies, sums, differences, partial permutations, clips about zero, sign rounding, and Γ gates, or a query block with coordinates $\Gamma_1(u_i, \ell)$, where $\|u(x)\|_2 \le O_{C,R,Z}(1)\|x\|_2$ and the internal scalar satisfies $\ell \in [0, 1]$. Then the complete map satisfies $\|\psi(x)\|_2 \le O_{C,R,Z}(1)\|x\|_2$ for all real inputs and $\psi(0) = 0$, independently of p, t. The thought-interface construction below verifies these conditions. The modification W_a has its separate polynomial allowance in Lemma F.7.

Proof. (i) If $\chi = 1$, the expression is $\sigma(v) - \sigma(-v) = v$. If $\chi = 0$, both arguments $v - M_0$ and $-v - M_0$ are nonpositive because $|v| \le M_0$, so both ReLUs vanish. (ii) For $x \le \theta_0$, both arguments are nonpositive; for $x \ge \theta_1$, both are nonnegative and their difference is $(\theta_1 - \theta_0)/(\theta_1 - \theta_0) = 1$; in between, only the first ReLU is active and the map is linear with slope $(\theta_1 - \theta_0)^{-1}$; $f(0) = 0$ when $\theta_0 \ge 0$ because then both arguments are nonpositive at $x = 0$. (iii) A one-hot block is entrywise nonnegative, so σ acts as the identity on Πh. On $\{0, 1\}^k$, the sum $\sum_i b_i - (k-1)$ equals 1 when all registers are 1 and is at most 0 otherwise; similarly $b_1 + b_2 \in \{0, 1, 2\}$ is clipped to $\min\{b_1 + b_2, 1\}$ by the two ReLUs. For the folded gate, $k - \sum_i b_i$ equals 0 when $\chi = 1$ and is an integer ≥ 1 otherwise, in which case the bias shift is at least M_0 and both ReLUs of (i) vanish. (iv) For $t \ge 1/(2\sqrt{d})$, both ReLUs are active and the expression is $-1/\sqrt{d} + 2(t + \frac{1}{2\sqrt{d}}) - 2(t - \frac{1}{2\sqrt{d}}) = 1/\sqrt{d}$; for $t \le -1/(2\sqrt{d})$, both vanish and it is $-1/\sqrt{d}$; for $|t| < 1/(2\sqrt{d})$, it equals $2t$, whence $q(0) = 0$ and $|q_{\mathrm{sgn}}(t)| \le 2|t|$ on the band, while outside the band $|q_{\mathrm{sgn}}(t)| = 1/\sqrt{d} \le 2|t|$. If $\|w - a\|_\infty \le 1/(2\sqrt{d})$ for a sign embedding a, then each coordinate w_j lies in $[1/(2\sqrt{d}), 3/(2\sqrt{d})]$ when $a_j = +1/\sqrt{d}$ and in $[-3/(2\sqrt{d}), -1/(2\sqrt{d})]$ when $a_j = -1/\sqrt{d}$, so $q(w_j) = a_j$. (v) is (ii) with $(\theta_0, \theta_1) = (\frac{1}{4}, \frac{3}{4})$ in the variable t: since $\theta_1 - \theta_0 = \frac{1}{2}$, $f_{1/4,3/4}(t) = \sigma(2t - \frac{1}{2}) - \sigma(2t - \frac{3}{2})$, and $\theta_0 = \frac{1}{4} \ge 0$. (vi) is (i) with the constant c in place of v. (vii) The first bracket is (i) with gate $1 - \chi$ (bias shift $M_0\chi$) and the second is (vi); on $\chi = 0$ they give $v + 0$ and on $\chi = 1$ they give $0 + c$. For these resource bounds, represent each signed coordinate needed later by $\sigma(v), \sigma(-v)$ and carry it by identity until its last use. Each coordinatewise gadget uses $O(1)$ units and nonzeros per coordinate; a reduction or broadcast uses $O(p)$, and a partial permutation uses at most one per coordinate. Separate affine stages avoid forming the dense product of a broadcast and a reduction. Padding with signed identities preserves these orders. There are $O_C(1)$ stages, proving the depth, width, and support claims. Summing squared coefficients, including biases, gives $O_{C,R,Z}(p + t^{-2})$, hence the norm bound. For the growth bound, $|\Gamma_A(g, v)| \le 2|v|$ for all real g, v; for $\ell \in [0, 1]$, $\Gamma_1(0, \ell) = 0$ and the map in u is 2-Lipschitz, so $|\Gamma_1(u, \ell)| \le 2|u|$. Clips about zero are nonexpansive, sign rounding has growth bound $|q_{\mathrm{sgn}}(v)| \le 2|v|$, and partial permutations have norm at most one. For $0 \le \kappa \le 1$, $\Gamma_{|c_i|}(\kappa, c_i) = \kappa c_i$, so the guarded constant has norm at most $2\|c\|_2|\iota|$ for either guard. A bounded number of compositions, copies, and sums preserves a dimension-independent linear growth bound. Internal scalar ramps need not satisfy such a bound individually: only their clamped values enter the specified gates. □

B THE ENVIRONMENT AND THE TWO-SIDED SEPARATION

Raw Gaussian rewards are never fed to the Transformer (they are unbounded and break the token radii, the loss envelope and the covering-number arguments); instead the observed rewards are the *bounded, strictly monotonically transformed* rewards of Lemma B.3. All probability statements use this unconditional distribution. The walk of Definition B.4 counts both edge directions; the stopped thought problem of Definition E.3 counts only fresh selections. Theorem F.5 relates these budgets without conditioning the fresh-instance distribution on gap events.

B.1 ENVIRONMENT AND REWARD TRANSFORMATION

Definition B.1 (Search environment)**.** Fix integers $C \geq 2$, $H \geq 2$ and $\bar{\zeta} \in (0, 1/2)$, and write $r_i = H - i + 1$, $S = S_C = (C^{H+1} - 1)/(C - 1)$ and $L_C = \lfloor \log_C H \rfloor$. At each internal node u of the complete C-ary depth-H tree, independently draw B_u uniformly from $[C]$. The target follows these preferred children. For a node of depth $i - 1$, set

$$Y_{u,b} = \mathbf{1}\{b = B_u\} + \sigma_i \xi_{u,b}, \qquad \xi_{u,b} \overset{\text{iid}}{\sim} N(0, 1), \qquad b \in [C],$$

where $\sigma_i > 0$ gives total misranking probability ζ_i and

$$\zeta_i = \begin{cases} \bar{\zeta}, & r_i \leq L_C, \\ H^{-3} C^{L_C - r_i}, & r_i > L_C. \end{cases}$$

Let $F_{i,1}, F_{i,0}$ be the $N(1, \sigma_i^2)$ and $N(0, \sigma_i^2)$ CDFs. The observed rewards are

$$r_{u,b} = \psi_{i,C}(Y_{u,b}), \qquad \psi_{i,C} = 2\bar{F}_i - 1, \qquad \bar{F}_i = \{F_{i,1} + (C - 1)F_{i,0}\}/C.$$

This is the probability integral transform using the marginal mixture CDF, rescaled to $(-1, 1)$. Only the tree and observed rewards are presented. The target indicator is read only at the committed node. Every internal node, including off-path nodes, has one mean-one child and $C - 1$ mean-zero children. For $C = 2$, the child slots are $\{0, 1\}$; write $\mathcal{E}_H := \mathcal{E}_H^{(2)}$, $L_H := L_2$ and $\psi_i := \psi_{i,2}$. These are exactly the binary environment, schedule and reward transform used in the separation theorem.

Lemma B.2 (Noise level for a prescribed misranking probability)**.** *The misranking function is*

$$q_C(\sigma) = 1 - \mathbb{E}_{Z \sim N(0,1)} \Phi(Z + 1/\sigma)^{C-1}.$$

It is continuous and strictly increasing from 0 *to* $1 - 1/C$ *as* σ *ranges from* 0 *to* ∞*. Every* ζ_i *above therefore determines a unique noise level* σ_i*. Node misranking indicators are independent. Strictly increasing depth transforms preserve every sibling order. For* $C = 2$*,* $q_2(\sigma) = \bar{\Phi}(1/(\sqrt{2}\sigma))$*, so* $\sigma_i = [\sqrt{2}\, \bar{\Phi}^{-1}(\zeta_i)]^{-1}$.

Proof. Condition on the preferred child's noise Z. Each other noise is below $Z + 1/\sigma$ independently, which gives the formula. For each finite Z, $\Phi(Z + 1/\sigma)^{C-1}$ decreases strictly with σ; bounded convergence proves continuity and both endpoint limits. At infinity, $\Phi(Z)$ is uniform on $(0, 1)$, so its $(C - 1)$st moment is $1/C$. The node blocks are independent. The binary formula follows by subtracting two independent unit Gaussians. All schedule probabilities lie in $(0, 1/2) \subseteq (0, 1 - 1/C)$, with an open upper endpoint also for $C = 2$. □

Lemma B.3 (Bounded reward transformation and gap bounds)**.** *(i) The map* $\psi_{i,C}$ *is a strictly increasing bijection from* $\mathbb{R}$ *to* $(-1, 1)$ *and preserves the observation sigma-algebra. (ii) Rewards have radius* $R = 1$*. (iii) Conditional on* B_u*, the preferred and non-preferred densities obey* $g_1 \leq C/2$*,* $g_0 \leq C/[2(C - 1)]$*. For ordered rewards* $r^u_{(1)} > \cdots > r^u_{(C)}$ *put* $\Gamma_{u,j} = r^u_{(j)} - r^u_{(j+1)}$*. For every* $t > 0$*,*

$$\mathbb{P}(\min_{j<C} \Gamma_{u,j} < t) \leq \mathbb{E} \sum_{j<C} \mathbf{1}\{\Gamma_{u,j} < t\} \leq \kappa_C t, \qquad \kappa_C = \binom{C}{2}.$$

For any fixed subset of children, its top-two gap Δ *satisfies* $\mathbb{P}(\Delta < t) \leq \min\{1, Ct/2\}$*, with* $\Delta = +\infty$ *for at most one child. All bounds also hold conditionally on any value of* B_u*.*

Proof. Suppress the depth index. With $y = \psi_{i,C}^{-1}(z)$ and Gaussian densities f_a, the transformed densities are

$$g_a(z) = \frac{C f_a(y)}{2\{f_1(y) + (C-1) f_0(y)\}}, \qquad g_1 + (C-1) g_0 = C/2.$$

The likelihood ratio $f_1/f_0 = \exp((2y-1)/(2\sigma_i^2))$ increases, so g_0 decreases and $g_0(1) = 0$ by a limit. This proves the density bounds; the positive derivative and endpoint limits of $\psi_{i,C}$ prove (i),(ii). Conditional on B_u, all child rewards are independent.

All gaps. Translate $(-1, 1)$ to $(0, 2)$ and write $q = g_0$. Then q decreases, $0 \le q \le 1$, and $\int_0^2 q = 1$. For $0 \le v \le 1$, set $l = 2 - v$, $A = \int_0^l q(x)\,dx$ and $D = \int_0^l q(x+v)\,dx$. Monotonicity gives $A \ge l/2 \ge D$; normalization and $q \le 1$ give $A \le 1$, $D \ge 1 - v$. Two decreasing functions have nonnegative covariance (integrate $(q(x) - q(y))(q(x+v) - q(y+v)) \ge 0$), hence $\int_0^l q(x)q(x+v)\,dx \ge AD/l$. Therefore

$$A + D - 2\int_0^l q(x)q(x+v)\,dx \le \frac{l}{2} + \frac{2(A - l/2)(l/2 - D)}{l} \le \frac{l}{2} + \frac{v^2}{2l} \le 1.$$

Using $g_1 = C/2 - (C-1)q$, the density at spacing v of the expected number of unordered child pairs is exactly $\binom{C}{2}$ times the left side. Integrating over $0 < v < t \le 1$ bounds the expected pair count by $\binom{C}{2}t$. For $t > 1$, the total pair count is at most $\binom{C}{2}$. Every small adjacent gap is a distinct such pair, proving the event and count bounds. Ties have probability zero.

Top gap. Return to $(-1, 1)$ and let F, G be the CDFs of $g_1, q = g_0$. Stochastic dominance gives $F \le G$. A subset of $k \ge 2$ non-preferred children has top-gap density at most $k\|q\|_\infty \le C/2$: condition on the maximum of the other $k-1$ rewards and sum over the possible winner. For a subset with the preferred child and k others, $1 \le k \le C-1$, put $D_k := FG^{k-1} - (C-1)G^k/k = G^{k-1}\big(F - \frac{C-1}{k}G\big) \le G^k\big(1 - \frac{C-1}{k}\big) \le 0$ (for $k = 1$ this is $F - (C-1)G \le 0$, including $C = 2$). Conditioning on the runner-up at x gives its top-gap density at $v \in (0, 2)$ as

$$h(v) = \frac{C}{2}G(1-v)^k + k\int_{-1}^{1-v} q(x+v)D_k'(x)\,dx \le \frac{C}{2}.$$

Indeed, the runner-up is either preferred or non-preferred; expanding these two cases and substituting $g_1 = C/2 - (C-1)q$ gives the equality. Integrating by parts on $[-1, 1-v-\epsilon]$ and letting $\epsilon \downarrow 0$, $\int q(x+v)D_k'(x)\,dx = [q(x+v)D_k(x)]_{-1}^{1-v} - \int q'(x+v)D_k(x)\,dx \le 0$: the upper boundary term vanishes because $q(1) = 0$, the lower one because $F(-1) = G(-1) = 0$, and the remaining integrand satisfies $q'D_k \ge 0$. Integrating h proves the claim. The subset A must be fixed independently of this outgoing reward block; the bound also holds for a random subset A that is conditionally independent of $(Y_{u,b})_{b\in[C]}$ given B_u. □

B.2 DFS EDGE TRAVERSALS AND THE SEPARATION THEOREM

Definition B.4 (Edge traversals and the stopped DFS walk)**.** On the observed tree of Definition B.1, an edge traversal goes from a node to a child or its parent. A forward walk uses only parent-to-child traversals; a backtracking walk may also return to the parent. The reward-ordered DFS teacher selects the highest-reward unvisited child, and returns when none remains. Write $u_0, u_1, \ldots$ for its visited nodes and $\tau_G := \min\{k : u_k = s^\star\}$ for its first goal visit. With walk budget B, observe it through $\min\{\tau_G, B\}$ traversals. Thus the traversal count includes both selections and returns, whereas the thought-interface budget of Definition E.3 counts only fresh selections.

Lemma B.5 (Posterior factorization)**.** *Given the full observed field $\mathcal{O}$, the preferences remain independent. With $y_b = \psi_{i,C}^{-1}(r_{u,b})$, their marginals are*

$$p_{u,b} = \mathbb{P}(B_u = b \mid \mathcal{O}_u) = \frac{\exp(y_b/\sigma_i^2)}{\sum_{a=1}^{C}\exp(y_a/\sigma_i^2)}.$$

For a leaf ℓ, $\mathbb{P}(s^\star = \ell \mid \mathcal{O}) = \prod_{u\in\mathrm{path}(\ell)} p_{u,b_\ell(u)}$, and $\mathbb{E}[1 - \max_b p_{u,b}] = \zeta_i$.

Proof. The likelihood for $B_u = b$ is $f_{i,1}(y_b)\prod_{a\neq b} f_{i,0}(y_a)$ and $f_{i,1}(y)/f_{i,0}(y) = \exp((2y-1)/(2\sigma_i^2))$. Uniform priors give the display. Independent node blocks and a bijective observation transform give the conditional product law. The leaf formula is the definition of the target. The Bayes decision is the largest utility, so its Bayes error is exactly the prescribed total misranking probability. For $C = 2$, with slots indexed by $\{0,1\}$, set $p_u := p_{u,1}$. Then $p_{u,0} = 1 - p_u$, $\mathbb{E}\min\{p_u, 1-p_u\} = \zeta_i$, and $\pi(\ell \mid \mathcal{O}) := \mathbb{P}(s^\star = \ell \mid \mathcal{O}) = \prod_{u\in\mathrm{path}(\ell)} p_u^{b_\ell(u)}(1-p_u)^{1-b_\ell(u)}$. □

Binary separation setting. The next theorem specializes to $C = 2$, with $\mathcal{E}_H = \mathcal{E}_H^{(2)}$ and $L_H = L_2$. Its upper bound retains the binary posterior second moment.

Theorem B.6 (Single-path forward and DFS separation under full observation)**.** *In $\mathcal{E}_H(\bar\zeta)$ with $\bar\zeta \in (0, \frac{1}{2})$:*

(a) ***(Full-observation single-path success bound.)*** *Consider any procedure, possibly randomized, that may inspect the* entire *reward-labeled tree before and while acting, but whose trajectory moves only from parent to child and never revisits a node. Then, for every $q > 1$,*

$$\mathbb{P}(\textit{reach } s^\star) \;\le\; \prod_{i=1}^{H} \kappa_{i,q}^{1/q}, \qquad \kappa_{i,q} := \mathbb{E}\big[p_i^{\,q} + (1-p_i)^{\,q}\big],$$

with p_i the per-node posterior at depth $i-1$. With $q = 2$,

$$\mathbb{P}(\textit{reach } s^\star) \;\le\; \prod_{i=1}^{H} \kappa_{i,2}^{1/2} \;\le\; \kappa_2(\bar\zeta)^{\lfloor \log_2 H\rfloor/2} \;\le\; C'_{\bar\zeta}\, H^{-\gamma_{\mathrm{obs}}},$$

where the common bottom-level moment and constants are

$$\kappa_2(\bar\zeta) := \mathbb{E}[p^2 + (1-p)^2] < 1 - \bar\zeta, \quad \gamma_{\mathrm{obs}} := -\tfrac{1}{2}\log_2 \kappa_2(\bar\zeta), \quad C'_{\bar\zeta} := \kappa_2(\bar\zeta)^{-1/2}.$$

The bound is architecture-, training-, sample-, and computation-independent, and holds for any presentation that is a measurable function of $\mathcal{O}$ and instance-independent randomness (Definition D.1).

(b) ***(DFS budget guarantee.)*** *The reward-ordered DFS walk on the observed rewards satisfies $\mathbb{P}\big(\tau_G \le 5H\big) \ge 1 - H^{-3}$.*

Proof. (a) Fix the observation field $\mathcal{O}$ and the procedure's internal randomness. A parent-to-child, no-revisit trajectory selects one leaf, so the conditional success probability is at most $M_H := \max_\ell \pi(\ell \mid \mathcal{O})$. For a height-$h$ subtree with root posterior p and child-subtree maxima $M_{h-1}^{(0)}, M_{h-1}^{(1)}$ (independent of each other and of p by Lemma B.5 and disjoint observation blocks)

$$M_h = \max\big\{(1-p)M_{h-1}^{(0)},\; pM_{h-1}^{(1)}\big\}, \qquad M_h^q \le (1-p)^q\big(M_{h-1}^{(0)}\big)^q + p^q\big(M_{h-1}^{(1)}\big)^q.$$

Taking expectations and iterating over depth (with $M_0 = 1$ at leaves) gives $\mathbb{E}M_H^q \le \prod_i \kappa_{i,q}$, and $\mathbb{E}M_H \le (\mathbb{E}M_H^q)^{1/q}$ by Jensen. For $q = 2$, with $m := \min\{p, 1-p\}$, the finite nonzero Gaussian noise gives $0 < m < 1/2$ almost surely. Thus

$$\kappa_{i,2} = 1 - 2\mathbb{E}m + 2\mathbb{E}m^2 < 1 - \mathbb{E}m = 1 - \zeta_i$$

by Lemma B.5. Dropping the upper-level factors, which are at most one, leaves $\kappa_2(\bar\zeta)^{L_H/2}$. Since $L_H > \log_2 H - 1$, this is at most $\kappa_2(\bar\zeta)^{-1/2} H^{\frac{1}{2}\log_2 \kappa_2(\bar\zeta)}$.

(b) By Lemma B.3(i) the walk on observed rewards coincides with the walk on utilities. Reveal the instance along the walk: each first arrival at an internal node exposes a fresh, independent flip indicator (Lemma B.2). If the target-path node at depth $i-1$ is flipped, DFS exhausts the disjoint non-preferred sibling subtree and returns, adding exactly $2(2^{r_i}-1)$ edge traversals irrespective of the reward ordering inside it; otherwise it advances. The walk identity and its tail bound (Lemma B.7) at $\tau = L_H$ give $\mathbb{P}(\tau_G > H + 2^{L_H+2}) \le Z_{>L_H} = \sum_{r>L_H} H^{-3} 2^{\,L_H - r} < H^{-3}$, and $2^{L_H} \le H$ makes $H + 2^{L_H+2} \le 5H$. □

B.3 THE DEPTH-FIRST WALK

We return to arbitrary $C \geq 2$, with $r_i = H - i + 1$ and $Z_{>\tau}(\boldsymbol{\zeta}) = \sum_{i:r_i>\tau} \zeta_i$.

Lemma B.7 (Walk identity and tail)**.** *Let $J_i \in \{1, \dots, C\}$ be the observed rank of the preferred child at the target-path node of depth $i-1$. These ranks are independent and $\mathbb{P}(J_i > 1) = \zeta_i$. On every tie-free instance,*

$$\tau_G = H + 2\sum_{i=1}^{H}(J_i - 1)\frac{C^{r_i} - 1}{C - 1}.$$

For any integer $0 \leq \tau \leq H$, let

$$D_C(\tau) = H + 2\left\{\frac{C(C^\tau - 1)}{C - 1} - \tau\right\}, \qquad Z_{>\tau} = \sum_{r_i>\tau} \zeta_i.$$

Then $\mathbb{P}(\tau_G > D_C(\tau)) \leq Z_{>\tau}$, and

$$\frac{2}{C-1}\sum_i \zeta_i(C^{r_i} - 1) \leq \mathbb{E}\tau_G - H \leq 2\sum_i \zeta_i(C^{r_i} - 1).$$

Proof. Condition on all preferences, fixing the target path. Disjoint outgoing noise blocks give independent ranks whose laws depend only on depth, not on the conditioned preferences. This proves independence also without conditioning. Before advancing along the preferred edge at level i, DFS exhausts $J_i - 1$ wrong sibling subtrees. Each contains $(C^{r_i} - 1)/(C - 1)$ entering edges, including its edge from the path, each traversed once each way. These subtrees are disjoint. The H path edges are traversed once before reaching the target, proving the identity. Also $\zeta_i \leq \mathbb{E}(J_i - 1) \leq (C - 1)\zeta_i$, proving the mean bounds. If no rank exceeds one at levels with $r_i > \tau$, the excess walk length is at most $2\sum_{r=1}^{\tau}(C^r - 1) = D_C(\tau) - H$. A union bound proves the tail. For $C = 2$, the bounds also give the weaker tail $\mathbb{P}(\tau_G > H + 2^{\tau+2}) \leq Z_{>\tau}$ as a specialization. Here, $F_i = J_i - 1$ is the binary flip indicator and $\mathbb{E}\tau_G = H + 2\sum_i \zeta_i(2^{r_i} - 1)$. □

Lemma B.8 (C-ary DFS budget)**.** *The noise schedule in Definition B.1 satisfies*

$$Z_{>L_C} = \frac{H^{-3}}{C-1}\left(1 - C^{L_C - H}\right) \leq \frac{H^{-3}}{C-1} \leq H^{-3}.$$

Reward-ordered DFS satisfies $\mathbb{P}(\tau_G \leq D_C(L_C)) \geq 1 - Z_{>L_C}$, where $D_C(L_C) = H + 2C(C^{L_C} - 1)/(C-1) - 2L_C \leq 5H$. In particular, $\mathbb{P}(\tau_G \leq 5H) \geq 1 - H^{-3}$.

Proof. The geometric sum over $L_C < r \leq H$ gives the displayed mass; its infinite extension is $H^{-3}/(C-1)$. Except on an event of probability at most $Z_{>L_C}$, every preferred child at levels with $r_i > L_C$ has rank one. Since $J_i - 1 \leq C - 1$, the walk identity on this event gives

$$\begin{aligned}\tau_G &\leq H + 2\sum_{r=1}^{L_C}(C^r - 1) = D_C(L_C) \\ &\leq H + \frac{2C(C^{L_C} - 1)}{C-1} \leq H + 4C^{L_C} \leq 5H.\end{aligned}$$

Here, $C/(C-1) \leq 2$ and $C^{L_C} \leq H$. Empty sums at $L_C = 0$ are zero. No restriction on ranks within the suffix is needed. □

Remark B.9 (Scope of the depth-first witness). Reward-ordered DFS reaches the goal within $5H$ edge traversals with probability at least $1 - H^{-3}$ for every $C \geq 2$ (Lemma B.8); its $C = 2$ specialization is Theorem B.6(b). No optimality among backtracking policies is claimed.

C COVERING NUMBERS AND DECODING UNDER BOUNDED ERRORS

This appendix gives output envelopes and covering-number bounds for the norm-bounded causal classes. These bounds control the number of on-policy training rounds in Appendix E and give guarantees from teacher-forced prediction errors in Appendix F.

Conventions. Fix the architecture and norm bounds $\mathbf{\Lambda} = (L_{\mathrm{att}}, J, L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_V, \Lambda_B, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$ and write $p = p_{\mathrm{TI}}$. An FFN of $\Psi_p(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$ is a map

$$\psi(x) = W_L\,\sigma\big(W_{L-1}\,\sigma(\cdots\sigma(W_1x + b_1)\cdots) + b_{L-1}\big) + b_L, \qquad L = L_{\mathrm{ffn}},$$

with σ the coordinatewise ReLU, $W_1 \in \mathbb{R}^{W_{\mathrm{ffn}}\times p}$, $W_i \in \mathbb{R}^{W_{\mathrm{ffn}}\times W_{\mathrm{ffn}}}$ for $1 < i < L$, $W_L \in \mathbb{R}^{p\times W_{\mathrm{ffn}}}$, every $\|W_i\|_{\mathrm{op}} \le \Lambda_{\mathrm{ffn}}$, every $\|b_i\|_2 \le \Lambda_{\mathrm{ffn}}$, and the linear growth bound $\|\psi(x)\|_2 \le c_{\mathrm{ffn}}\|x\|_2$ for all x (for $L = 1$ the map is affine; a bias-free layer is the case $b_i = 0$). Without loss of generality $c_{\mathrm{ffn}} \ge 1$ and $\Lambda_{\mathrm{ffn}} \ge 1$: a class with smaller constants is contained in the class with these. Let B_x bound the column norm of the input seen by the trainable Transformer (for the thought interface including the interface-written records, and for the causal class $\mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}$ of Definition A.2 including the concatenated positional block), and let B_y bound the norm of every target token. For a token matrix $X = [x_1, \ldots, x_m]$, write $\|X\|_{2,\infty} := \max_i \|x_i\|_2$. The trainable parameters of $F \in \mathcal{F}_{\mathrm{TI}}(\mathbf{\Lambda})$ are the $L_{\mathrm{att}}J$ matrix pairs (V_h, B_h) and the $L_{\mathrm{att}} + 1$ FFN tuples $(W_{r,i}, b_{r,i})_{0\le r\le L_{\mathrm{att}},\, i\le L_{\mathrm{ffn}}}$; we write θ for the full parameter vector, F_θ for the Transformer it defines, and

$$\|\theta - \theta'\|_{\mathrm{par}} := \max\big\{\|V_h - V'_h\|_{\mathrm{op}},\ \|B_h - B'_h\|_{\mathrm{op}},\ \|W_i - W'_i\|_{\mathrm{op}},\ \|b_i - b'_i\|_2\big\}$$

for the maximum over all parameter blocks. Let $D_{\mathrm{TI}} = D_{\mathrm{TI}}(\mathbf{\Lambda})$ be the number of scalar entries of θ,

$$D_{\mathrm{TI}} = O\big(L_{\mathrm{att}}Jp^2 + (L_{\mathrm{att}} + 1)L_{\mathrm{ffn}}(W^2_{\mathrm{ffn}} + pW_{\mathrm{ffn}} + p)\big);$$

for the thought interface the fixed front end is not counted. Every scalar entry of θ is bounded in absolute value by $\Lambda_{\max} := \max\{\Lambda_V, \Lambda_B, \Lambda_{\mathrm{ffn}}\}$. The per-instance loss of F is

$$\ell_F(\mathcal{T}) := \frac{1}{2}\sum_{k=0}^{L_{\mathrm{TI}}(\mathcal{T})-1} \big\|F(E_{\mathrm{TI},k}(\mathcal{T}))_{:,-1} - o_{\mathrm{TI},k+1}(\mathcal{T})\big\|_2^2,$$

so that $\mathcal{R}_{\mathrm{TI}}(F) = \mathbb{E}_{\mathcal{T}}\,\ell_F(\mathcal{T})$ and $\widehat{\mathcal{R}}_{\mathrm{TI},N}(F) = \mathbb{P}_N\ell_F$ with $\mathbb{P}_N$ the empirical measure of the N training instances. The covering number $\mathcal{N}(\alpha, \mathcal{F}, d_\infty)$ (van der Vaart & Wellner, 1996) is the smallest cardinality of an α-cover $\mathcal{G} \subseteq \mathcal{F}$: for every $F \in \mathcal{F}$, some $G \in \mathcal{G}$ satisfies $d_\infty(F, G) \le \alpha$. Here, $d_\infty(F, G) := \sup_E \|F(E)_{:,-1} - G(E)_{:,-1}\|_2$ is the supremum pseudometric on the declared bounded-prefix domain.

Lemma C.1 (Transformer output envelope)**.** *Let $B_\circ := c_{\mathrm{ffn}}^{L_{\mathrm{att}}+1}(1 + J\Lambda_V)^{L_{\mathrm{att}}}B_x$. For every $F \in \mathcal{F}_{\mathrm{TI}}(\mathbf{\Lambda})$ (or $\mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}$) and every prefix E with $\|E\|_{2,\infty} \le B_x$, the input of every attention layer and of every FFN of F on E has column norms at most $B_\circ$, and in particular*

$$\|F(E)_{:,-1}\|_2 \le B_{\mathrm{out}} := B_\circ.$$

Consequently, the per-instance loss satisfies

$$0 \le \ell_F(\mathcal{T}) \le \mathsf{B}_{\mathrm{TI}} := \frac{L_{\max,\mathrm{TI}}}{2}\big(B_{\mathrm{out}} + B_y\big)^2 \qquad \textit{for every instance } \mathcal{T}.$$

Proof. A residual attention layer maps a matrix with column norms at most B to a matrix with column norms at most $(1+J\Lambda_V)B$: the output of head h at column i is $V_hX\omega_i$ with ω_i a probability vector (for the causal class, supported on columns $\le i$), so $\|V_hX\omega_i\|_2 \le \Lambda_V \max_{i'}\|x_{i'}\|_2 \le \Lambda_V B$, and the residual term contributes $\|x_i\|_2 \le B$. An FFN maps column norms $\le B$ to column norms $\le c_{\mathrm{ffn}}B$ by the linear growth bound. The Transformer alternates $L_{\mathrm{att}} + 1$ FFNs and L_{att} attention layers starting from column norms $\le B_x$, and both factors are at least one, so every intermediate column norm is at most $B_\circ$. The loss bound follows from the triangle inequality, $\|F(E_{\mathrm{TI},k})_{:,-1} - o_{\mathrm{TI},k+1}\|_2 \le B_{\mathrm{out}} + B_y$, and $L_{\mathrm{TI}}(\mathcal{T}) \le L_{\max,\mathrm{TI}}$. □

Lemma C.2 (Softmax is Lipschitz from ℓ_∞ to ℓ_1)**.** *For score vectors $s, s' \in \mathbb{R}^m$, $\|\mathrm{softmax}(s) - \mathrm{softmax}(s')\|_1 \le \|s - s'\|_\infty$. The constant is sharp. The same bound holds for the softmax restricted to any index subset (as in the causal class), applied to the restricted score vectors.*

Proof. The Jacobian of softmax at s is $\mathrm{diag}(w) - ww^\top$ with $w = \mathrm{softmax}(s)$ (Gao & Pavel, 2017, Proposition 5). For $v \in \mathbb{R}^m$, $\|(\mathrm{diag}(w) - ww^\top)v\|_1 = \sum_i w_i\,|v_i - \langle w, v\rangle| \le \sqrt{\sum_i w_i v_i^2 - \langle w, v\rangle^2} \le \|v\|_\infty$ by Cauchy–Schwarz. Integrating along the segment from s to s' gives the claim; a restricted softmax is the softmax of the restricted vector. Equal binary weights with $v = (1, -1)$ attain the infinitesimal ratio one. □

Lemma C.3 (Parameter cover of the Transformer prediction class)**.** *Let $B_\circ$ be as in Lemma C.1 and set*

$$\ell_{\mathrm{att}} := 1 + J\Lambda_V\big(1 + 2\Lambda_B B_\circ^2\big), \qquad \pi_{\mathrm{att}} := JB_\circ\big(1 + \Lambda_V B_\circ^2\big),$$
$$\ell_{\mathrm{ffn}} := \Lambda_{\mathrm{ffn}}^{L_{\mathrm{ffn}}}, \qquad \pi_{\mathrm{ffn}} := L_{\mathrm{ffn}}\Lambda_{\mathrm{ffn}}^{L_{\mathrm{ffn}}}\big(B_\circ + L_{\mathrm{ffn}}\big),$$

$\ell_{\max} := \max\{\ell_{\mathrm{att}}, \ell_{\mathrm{ffn}}\}$, $\pi_{\max} := \max\{\pi_{\mathrm{att}}, \pi_{\mathrm{ffn}}\}$, *and*

$$A_F := (2L_{\mathrm{att}} + 1)\,\ell_{\max}^{2L_{\mathrm{att}}}\,\pi_{\max}, \qquad A_{\mathrm{pred}} := 4\sqrt{\max\{p, W_{\mathrm{ffn}}\}}\,\Lambda_{\max}\,A_F.$$

Then:

(i) (Parameter Lipschitz bound.) *For all parameter vectors θ, θ' of the class and every prefix E with $\|E\|_{2,\infty} \le B_x$,*

$$\|F_\theta(E) - F_{\theta'}(E)\|_{2,\infty} \le A_F\,\|\theta - \theta'\|_{\mathrm{par}}.$$

(ii) (Covering numbers.) *For every $\alpha > 0$,*

$$\log \mathcal{N}(\alpha, \mathcal{F}_{\mathrm{TI}}(\mathbf{\Lambda}), d_\infty) \le D_{\mathrm{TI}} \log\Big(1 + \frac{A_{\mathrm{pred}}}{\alpha}\Big).$$

A_{pred} is polynomial in p, W_{ffn}, B_x, and the norm radii (with exponents depending only on L_{att} and L_{ffn}).

(iii) (Scope.) *The bounds hold verbatim for the causal class $\mathcal{F}_{\mathrm{TI}}^{\mathrm{c}}$ (causal masking is a fixed restriction of every softmax to the columns $\le i$, and the positional blocks are fixed inputs accounted for in B_x), and for the thought interface, where the interface-written records only change the input domain of the trainable part and add no parameter. Any subclass whose parameters are a Lipschitz image of a* bounded *lower-dimensional parameter set (Corollary F.4, with D_{TI} replaced by D_{sp}; the source radius and the Lipschitz constant enter the cover constant A) satisfies (ii) with that dimension.*

Proof. Attention layer. Let X, X' have column norms at most $B_\circ$, let (V_h, B_h) and (V_h', B_h') be two parameter sets of the layer with $\delta_X := \|X - X'\|_{2,\infty}$, $\delta_V := \max_h \|V_h - V_h'\|_{\mathrm{op}}$, $\delta_{\mathrm{sc}} := \max_h \|B_h - B_h'\|_{\mathrm{op}}$. For a query column i and head h, the scores are $s_{ij} = x_j^\top B_h x_i$ and $s_{ij}' = x_j'^\top B_h' x_i'$, and

$$|s_{ij} - s_{ij}'| \le |(x_j - x_j')^\top B_h x_i| + |x_j'^\top (B_h - B_h') x_i| + |x_j'^\top B_h' (x_i - x_i')| \le 2\Lambda_B B_\circ\,\delta_X + B_\circ^2\,\delta_{\mathrm{sc}}.$$

By Lemma C.2, the attention weights satisfy $\|\omega_i - \omega_i'\|_1 \le 2\Lambda_B B_\circ \delta_X + B_\circ^2 \delta_{\mathrm{sc}}$, also under causal masking. The head outputs differ by

$$\begin{aligned}\|V_h X\omega_i - V_h' X'\omega_i'\|_2 &\le \|(V_h - V_h')X\omega_i\|_2 + \|V_h'(X - X')\omega_i\|_2 + \|V_h' X'(\omega_i - \omega_i')\|_2 \\ &\le B_\circ \delta_V + \Lambda_V \delta_X + \Lambda_V B_\circ \|\omega_i - \omega_i'\|_1.\end{aligned}$$

Adding the residual term and summing over the J heads, the layer outputs differ in $\|\cdot\|_{2,\infty}$ by at most

$$\big(1 + J\Lambda_V(1 + 2\Lambda_B B_\circ^2)\big)\delta_X + JB_\circ\delta_V + J\Lambda_V B_\circ^3 \delta_{\mathrm{sc}} \le \ell_{\mathrm{att}}\,\delta_X + \pi_{\mathrm{att}}\max\{\delta_V, \delta_{\mathrm{sc}}\}.$$

FFN. For inputs x, x' with norms at most $B_\circ$ and parameter tuples (W_i, b_i), (W_i', b_i') at distance $d_{\mathrm{par}} := \max_i \max\{\|W_i - W_i'\|_{\mathrm{op}}, \|b_i - b_i'\|_2\}$, write h_i for the i-th hidden activation of ψ on x ($h_0 = x$, $h_i = \sigma(W_i h_{i-1} + b_i)$ for $i < L$). Since the ReLU is 1-Lipschitz and $\|h_i\|_2 \le \Lambda_{\mathrm{ffn}}\|h_{i-1}\|_2 + \Lambda_{\mathrm{ffn}}$, induction gives $\|h_i\|_2 \le \Lambda_{\mathrm{ffn}}^i(B_\circ + i)$. Telescoping the composition,

$$\begin{aligned}\|\psi(x) - \psi'(x)\|_2 &\le \sum_{i=1}^{L} \Lambda_{\mathrm{ffn}}^{L-i}\big(\|W_i - W_i'\|_{\mathrm{op}}\|h_{i-1}\|_2 + \|b_i - b_i'\|_2\big) \\ &\le L\Lambda_{\mathrm{ffn}}^L(B_\circ + L)d_{\mathrm{par}} = \pi_{\mathrm{ffn}}\,d_{\mathrm{par}},\end{aligned}$$

and $\|\psi(x) - \psi(x')\|_2 \le \Lambda_{\mathrm{ffn}}^L\|x - x'\|_2 = \ell_{\mathrm{ffn}}\|x - x'\|_2$; both bounds apply columnwise to token matrices in $\|\cdot\|_{2,\infty}$.

Composition. By Lemma C.1, along both F_θ and $F_{\theta'}$, every stage input has column norms at most $B_\circ$, so the bounds above apply at every one of the $2L_{\mathrm{att}}+1$ stages. Writing $F_\theta - F_{\theta'}$ as a telescoping

sum over stages (in the r-th term, the two Transformers share the stages before r, differ in the parameters of stage r on a common input, and the remaining stages of $F_{\theta'}$ are applied to both outputs), the r-th term is at most $\ell_{\max}^{2L_{\mathrm{att}}+1-r}\pi_{\max}\|\theta-\theta'\|_{\mathrm{par}} \le \ell_{\max}^{2L_{\mathrm{att}}}\pi_{\max}\|\theta-\theta'\|_{\mathrm{par}}$ (each of the $2L_{\mathrm{att}}+1-r$ remaining stages contributes a factor at most $\ell_{\max} \ge 1$), and there are $2L_{\mathrm{att}}+1$ terms, proving (i).

Covers. A matrix block $M \in \mathbb{R}^{a\times b}$ with $\|M\|_{\mathrm{op}} \le \Lambda_{\max}$ has Frobenius norm at most $\sqrt{\min\{a,b\}}\,\Lambda_{\max}$, and a Frobenius $\varepsilon_{\mathrm{cov}}$-net of that Frobenius ball has at most $(1+2\sqrt{\min\{a,b\}}\Lambda_{\max}/\varepsilon_{\mathrm{cov}})^{ab}$ elements by the volumetric bound for Euclidean balls (Vershynin, 2018, Section 4.2); Frobenius $\varepsilon_{\mathrm{cov}}$-closeness implies operator-norm $\varepsilon_{\mathrm{cov}}$-closeness. Bias blocks are covered in the same way. Taking the product over all parameter blocks gives an $\varepsilon_{\mathrm{cov}}$-net of the parameter set in $\|\cdot\|_{\mathrm{par}}$ of log-cardinality at most $D_{\mathrm{TI}}\log(1+2\sqrt{\max\{p,W_{\mathrm{ffn}}\}}\Lambda_{\max}/\varepsilon_{\mathrm{cov}})$. Its elements need not satisfy the linear growth constraint; the standard replacement of every net element that lies within $\varepsilon_{\mathrm{cov}}$ of the feasible set by a feasible point within $\varepsilon_{\mathrm{cov}}$ of it turns this into a $2\varepsilon_{\mathrm{cov}}$-net of the feasible set with net elements inside the class. By (i), the images of a feasible $2\varepsilon_{\mathrm{cov}}$-net under $\theta \mapsto (E \mapsto F_\theta(E)_{:,-1})$ form a $2\varepsilon_{\mathrm{cov}}A_F$-cover in d_∞. Choosing $\varepsilon_{\mathrm{cov}} = \alpha/(2A_F)$ proves (ii) with the displayed A_{pred}.

Scope. For the causal class, the score computation, the softmax Lipschitz bound (restricted to columns $\le i$), and the value transport are unchanged, and the positional block enters only through B_x; for the thought interface, the trainable Transformer receives the prefix with its interface-written records, whose columns are bounded by B_x by definition, so the covers are covers of the composed maps. For a parametrization $\vartheta \mapsto \theta(\vartheta)$ that is L_θ-Lipschitz from a bounded set of dimension D_{sp} into $\|\cdot\|_{\mathrm{par}}$ (Corollary F.4), a net at scale α/L_θ maps to a net at scale α. Thus (ii) holds with D_{sp} and with L_θ in the cover constant. □

C.1 Exact decoding under bounded errors

Lemma C.4 (Exact decoding under bounded errors)**.** *(a) If a finite set $\mathcal{C}$ has minimum Euclidean separation at least one, every point at distance less than $1/2$ from $c \in \mathcal{C}$ has c as its unique nearest element. This applies to the discrete blocks of Definition E.1, including nullable embeddings and pointers. For the nonnullable node dictionary, the decoding radius is $\rho = \frac{1}{2}\min_{u\neq v}\|a_u - a_v\|_2$, with $\sqrt{(1-\varepsilon)/2} \le \rho \le \sqrt{(1+\varepsilon)/2}$ for $\varepsilon < 1/2$.*

(b) Under the quantized front end and full write-back, a prediction at joint distance $r < \min\{1/2,\rho\}$ from a teacher selected-state target is stored as that exact target, including terminal targets. Own positions are supplied by the interface and excluded from this distance.

Proof. For $c' \neq c$, the triangle inequality gives $\|z-c'\|_2 \ge 1-\|z-c\|_2 > 1/2 > \|z-c\|_2$. Distinct sign embeddings have squared distance $2-2\langle a_u, a_v\rangle \in [2-2\varepsilon, 2+2\varepsilon]$; a unit embedding or one-hot has distance one from the null value, and distinct bits have distance one. This proves (a). In (b), each discrete block is within $r < 1/2$ of its target and decodes exactly. At nonterminal teacher states, both thoughts are node embeddings; clipping to any radius $Z \ge 1$ cannot increase their distance to those embeddings, and distance $r < \rho$ makes rounding exact. At termination, the protocol zeros both thoughts. It resets inactive V, r, h and writes all decoded registers, so every predicted field is the teacher's. Supplied own positions complete the token. □

D The interface separation: assembly and remarks

This appendix specializes to $C = 2$ and assembles the posterior single-path upper bound and the teacher's $5H$ DFS walk. The single-path bound permits every observation presentation, including the complete reward-labeled tree, and arbitrary computation. The learned witness of Theorem F.5 instead uses local front-end records and $O_C(H)$ context. The definition and inclusion lemma below also cover interleaved computation.

D.1 Single-path procedures and their success bound

Definition D.1 (Single-path forward procedures)**.** On $\mathcal{E}_H(\bar{\zeta})$, a presentation of the observations is a prefix $E_{\mathrm{pres}} = \mathrm{pres}(\mathcal{O}, U)$ that is a measurable function of the observed reward-labeled tree $\mathcal{O}$

and of randomness U independent of the instance. It may reveal the complete reward-labeled tree, but may not read the latent bits or the target. Under *any* such presentation, at each commitment, the legal set consists of the selected-state tokens of the current node's children (at a leaf, the terminal token). A single-path forward procedure is any measurable map from prefixes to $\mathbb{R}^p$, realized by a Transformer or not, trained on any data by any method, or not trained at all (its parameters must be independent of the fresh test instance given its presentation), possibly with internal randomness independent of the instance, executed autoregressively with actions decoded at the environment boundary to the legal set, for an arbitrary step budget. The emitted token itself is not canonicalized. *Success* means the decoded action reaches $s^\star$ at some step.

Lemma D.2 (Single-path procedures satisfy the success bound)**.** *Every single-path forward procedure induces a parent-to-child, no-revisit trajectory whose leaf choice is measurable with respect to the instance observations available in its prompt together with independent internal randomness. Hence Theorem B.6(a) applies to it under every presentation: its success probability is at most* $C'_{\bar{\zeta}} H^{-\gamma_{\mathrm{obs}}}$.

Proof. Decoding each action into the legal set ensures that it selects a child of the current node. At a leaf, only the terminal action is legal, so at most H moves are informative and the trajectory selects one leaf. By Definition D.1, whatever the presentation reveals is a function of the observation field and of instance-independent randomness, and the procedure's parameters are independent of the instance given the presentation; conditionally on $\mathcal{O}$, the leaf choice is therefore a randomized function of $\mathcal{O}$ alone, and its conditional success probability is at most $\max_\ell \pi(\ell \mid \mathcal{O})$, the quantity bounded in Theorem B.6(a). (Without the measurability requirement the claim fails: a presenter who knew $s^\star$ could encode it in the column order.) □

Assembly of Theorem 1. Lemma D.2 applies Theorem B.6(a) to every single-path forward procedure; Lemma B.8 at $C = 2$ supplies its DFS clause. The learned local-observation witness is Theorem F.5, whose budget counts fresh selections and whose context is $O_C(H)$.

Corollary D.3 (The single-path bound permits interleaved computation)**.** *Let a procedure interleave arbitrary tokens ("thoughts") between its node commitments, subject only to the single-path constraint that the committed nodes form a single root-to-leaf path; observation, budget, architecture, and computation are unrestricted as in Theorem B.6(a). Then every bound of Theorem B.6(a) holds verbatim.*

Proof. The class of Theorem B.6(a) is defined by the node commitments alone and already allows arbitrary internal computation; interleaved tokens neither change the environment's observation field nor relax the single-commitment structure, so conditioning on the field and the procedure's randomness, the committed leaf is unchanged in law and the posterior-mass bound applies. □

E THE THOUGHT INTERFACE: EXECUTION AND ON-POLICY LEARNING

This appendix proves exact thought-level execution and stopped closed-loop DFS for fixed $C \geq 2$ under Assumption 1. It then proves on-policy learning with copy labels on raw prefixes under raw feedback and a stopped bound on accumulated copy errors. Appendix F gives the finite-cover sufficient condition and original-interface counterexample under the quantized front end on $\mathcal{E}_H^{(C)}$. We use the sign-node dictionary and FFN gadgets of Appendix A. TI positions are one-hot on the declared context of $T = (C+2)(B_{\mathrm{sel}}+1)$ columns (Definition E.3), so $\varepsilon_{\mathrm{pos}} = 0$. Node embeddings have unit norm and incoherence ε (the $\varepsilon_{\mathrm{emb}}$ of the main text), thoughts have norm at most Z, and rewards have magnitude at most R. The executor uses signed exact reads and FFN comparisons.

E.1 THE CAUSAL THOUGHT INTERFACE AND THE SEARCH RULE

Each visit, called an *occurrence* in this appendix, carries a completed-slot mask, valid and completed counts, an occurrence identifier, the parent's occurrence identifier, and a backlink to the parent anchor. These registers store the search state in emitted tokens. In this appendix, *fresh* abbreviates a new visit created by a selection, including a visit whose node embedding appeared earlier; a return instead resumes an existing visit. Thus the fresh identifiers below identify visits, not node

embeddings. For branching factor C, C exact fixed-offset reads recover the current candidates or the parent's records according to the counter route. One addressed read recovers the linked parent anchor; a shared uniform head cancels the background of all $C+1$ reads. The output FFN compares rewards and matches the returning embedding to its parent slot. Every used address is visible. The interface writes the fixed front end's records and queries the Transformer only at anchors.

Definition E.1 (Causal thought interface)**.** Superscripts ph, slot, and pos distinguish phase, slot, and positional one-hots where they occur together below. For the layout and fixed offsets, index the C records by $j \in \{1, \dots, C\}$; record j corresponds to candidate slot $j-1 \in [C]$ in the main text. Tokens carry the blocks $(z, z_{\mathrm{src}}, \mathsf{P}, U, V, r, \phi, b, \nu, \iota, \mathsf{w}, n_{\mathrm{v}}, n_{\mathrm{c}}, \pi_{\uparrow}, \pi_{\mathrm{fr}}, \pi_{\mathrm{fr}\uparrow}; \pi; h)$: thought $z \in \mathbb{R}^d$ and source register z_{src}; parent and current-node embeddings P, U; candidate embedding V and reward r; phase one-hot $\phi \in \{e_1^{\mathrm{ph}}, \dots, e_{C+1}^{\mathrm{ph}}\}$; return flag $b \in \{0,1\}$; validity bit $\nu \in \{0,1\}$; constant $\iota = 1$; completed-slot mask $\mathsf{w} \in \{0,1\}^C$; counters $n_{\mathrm{v}}, n_{\mathrm{c}} \in \{e_0, \dots, e_C\}$; backlink $\pi_{\uparrow}$, fresh identifier π_{fr}, and the parent's fresh identifier $\pi_{\mathrm{fr}\uparrow}$, content blocks valued in the positional dictionary or zero; the column's own positional embedding π (interface metadata, Definition A.2); and scratch h, zero on all interface columns and targets. Given an initial thought $\|z_{\mathrm{root}}\|_2 \le Z$, the root token has $z = z_{\mathrm{src}} = z_{\mathrm{root}}$, $\mathsf{P} = a_0$, $U = a_{\mathrm{root}}$, $\phi = e_1^{\mathrm{ph}}$, $\iota = 1$, $n_{\mathrm{v}} = n_{\mathrm{c}} = e_0$, and every other predicted block zero; its own position is e_1^{pos}. For the quantized front end, $z_{\mathrm{root}} = a_{\mathrm{root}}$. The terminal token has $\mathsf{P} = a_0$, $U = a_{\mathrm{eos}}$, $\phi = e_1^{\mathrm{ph}}$, $\iota = 1$, $n_{\mathrm{v}} = n_{\mathrm{c}} = e_0$, and every other predicted block zero; no records follow it. A canonical nonterminal selected-state target has $z_{\mathrm{src}} = z$, $V = 0$, $r = 0$, $\nu = 0$, $\phi = e_1^{\mathrm{ph}}$; it is fresh if $b = 0$ and a return token if $b = 1$. A *candidate record* of slot j has $\phi = e_{j+1}^{\mathrm{ph}}$, $b = 0$, carries a proposal (z_j, a_{v_j}, r_j) in (z, V, r) with $\nu = \mathbf{1}\{a_{v_j} \neq a_{\mathrm{eos}}\}$, and copies of $z_{\mathrm{src}}, \mathsf{P}, U, \mathsf{w}, n_{\mathrm{c}}, \pi_{\uparrow}, \pi_{\mathrm{fr}\uparrow}$ of its occurrence. *Decoding and storage.* Let $\mathcal{A}$ be the full sign dictionary, including a_0, a_{eos}. For any finite set D, use nearest Euclidean decoding Q_D, with a fixed total order for ties, zero first when present. The map Dec acts on every discrete block:

block	decoding set
U	$\mathcal{A}$
P, V	$\mathcal{A} \cup \{0\}$
b, ν, w_j	$\{0,1\}$
ι	$\{1\}$
ϕ	$\{e_1^{\mathrm{ph}}, \dots, e_{C+1}^{\mathrm{ph}}\}$
$n_{\mathrm{v}}, n_{\mathrm{c}}$	$\{e_0, \dots, e_C\}$
$\pi_{\uparrow}, \pi_{\mathrm{fr}}, \pi_{\mathrm{fr}\uparrow}$	$\{0, e_1^{\mathrm{pos}}, \dots, e_T^{\mathrm{pos}}\}$.

Mask bits are decoded separately. Rewards, thoughts and scratch are not discrete blocks; own positions are exact metadata. Decoding enforces no parent–child legality. Registers are decoded correctly within distance $1/2$ of their prescribed values, as are nullable P, V and pointers. For nonnullable node embeddings, the decoding radius is $\rho_{\mathrm{code}} := \frac{1}{2} \min_{u \neq v} \|a_u - a_v\|_2 > 1/2$. Ties lie outside these open balls.

Under raw feedback, store the raw prediction after clipping only z, z_{src} to radius Z. Inactive V, r, h are retained. Read decoded discrete blocks to dispatch fresh selections and returns, recognize termination and acceptance, and count fresh selections with an exact interface counter. Stop at acceptance, termination, or the selection budget before appending further records. Malformed decoded actions have a fixed failure continuation. The analytical scope checks in Definition E.7 precede all actions and stops from the tested emission; they monitor errors without correcting feedback.

After a fresh state at column i, if execution continues, call

$$(z_j, a_{v_j}, r_j) := g_{\mathrm{pre}}(z, \mathrm{Dec}_U(U), e_{j-1}^{\mathrm{slot}}), \qquad j = 1, \dots, C,$$

with the interface's exact slot index. Each record copies the raw $z_{\mathrm{src}}, \mathsf{P}, U, \mathsf{w}, n_{\mathrm{c}}, \pi_{\uparrow}, \pi_{\mathrm{fr}\uparrow}$ once into the corresponding fields. It writes exact $\phi = e_{j+1}^{\mathrm{ph}}$, $b = 0$, $\iota = 1$, proposal z, V, r, validity ν, count $n_{\mathrm{v}} = e_n$ of all valid proposals, $\pi_{\mathrm{fr}} = e_i^{\mathrm{pos}}$, own position, and zero scratch. Valid record indices may form any subset of $\{1, \dots, C\}$, with pairwise distinct valid embeddings. Eligibility is $\nu_j(1 - \mathsf{w}_j)$; each returned eligible slot increments n_{c} once, so the total count detects exhaustion without any assumption on slot order. All C records are still written and read at fixed offsets. Fresh canonical state targets have $\pi_{\mathrm{fr}} = 0$ and $n_{\mathrm{v}} = e_0$. Returns receive no records. The *anchor* of an occurrence is a column at which a control decision is due: the C-th candidate record of a fresh occurrence, and

every return token. The Transformer's output is read at anchors only. The token width including exact-read scratch is $p_{\mathrm{TI}} = O(C(d+T+C))$, with $d_{\mathrm{pos}} = T$.

Definition E.2 (Search rule on the tree the rollout generated)**.** For any numerical prefix, define its decoded prefix $S(E)$ column by column. Apply Dec to discrete blocks, retain actual thoughts and exact metadata, and clear scratch. On decoded state columns, set inactive V, r to zero in $S(E)$ only; retain record rewards and proposals. Formally, $S(\varnothing) = \varnothing$ and $S(E \parallel x) = S(E) \parallel s(x)$ for this column map s; $S(\dagger) = \dagger$. This total map uses no search-rule evaluation. Raw record copies commute with decoding of the discrete blocks, and decoded-address front-end calls agree with replay of $S(E)$.

A canonical prefix is well-formed if it consists of the root and its prescribed records, then anchor outputs and protocol-written records, with legal discrete blocks, phases, links, masks, counts, visible references, and thoughts bounded by Z. The two learner-state thoughts need not coincide: the source is the actual clipped register copied into records. Canonical here refers to discrete blocks and prescribed inactive zeros, not to thoughts. Numerical admission is defined recursively in Definition E.7, before any label is evaluated. For a well-formed prefix E ending at an anchor x, the records of the current occurrence are the candidate records carrying $\pi_{\mathrm{fr}}(x)$, and such a record is eligible if it is valid and its slot is not marked in $\mathsf{w}(x)$. The correct next token $\mathsf{H}_{\mathrm{C}}(E)$ is:

(S) selection, if $n_{\mathrm{c}}(x) < n_{\mathrm{v}}(x)$: with $i^\star$ the eligible record of largest reward, the fresh selected-state token with $z = z_{\mathrm{src}} = z_{i^\star}$, $\mathsf{P} = U(x)$, $U = a_{v_{i^\star}}$, $b = 0$, $\mathsf{w} = 0$, $n_{\mathrm{v}} = n_{\mathrm{c}} = e_0$, $\pi_\uparrow = \pi(x)$, $\pi_{\mathrm{fr}} = 0$, $\pi_{\mathrm{fr}\uparrow} = \pi_{\mathrm{fr}}(x)$;

(R) return, otherwise if $\mathsf{P}(x) \neq a_0$: with y the column whose positional embedding is $\pi_\uparrow(x)$ (the parent's anchor) and j the slot of the record carrying $\pi_{\mathrm{fr}\uparrow}(x)$ whose candidate embedding is $U(x)$, the return token with $z = z_{\mathrm{src}} = z_{\mathrm{src}}(y)$, $\mathsf{P} = \mathsf{P}(y)$, $U = U(y)$, $b = 1$, $\mathsf{w} = \mathsf{w}(y) \vee e_j$, $n_{\mathrm{v}} = n_{\mathrm{v}}(y)$, $n_{\mathrm{c}} = \mathrm{shift}(n_{\mathrm{c}}(y))$ (one-hot shift by one), $\pi_\uparrow = \pi_\uparrow(y)$, $\pi_{\mathrm{fr}} = \pi_{\mathrm{fr}}(y)$, $\pi_{\mathrm{fr}\uparrow} = \pi_{\mathrm{fr}\uparrow}(y)$;

(T) termination, otherwise: emit the terminal token of Definition E.1.

The rule is a function of the prefix and of g_{pre} alone. The tree the rollout generated has one vertex per fresh occurrence, carrying the records the interface wrote from the thought the rollout actually emitted there, and one edge per selection. A rollout is closed-loop correct through step k if at every anchor $i \le k$ before stopping, its emission $\widehat{o}_i$ satisfies

$$\mathrm{Dec}(\widehat{o}_i) = \mathrm{Dec}\big(\mathsf{H}_{\mathrm{C}}(S(E_{i-1}))\big)$$

on all discrete blocks. Thoughts are not constrained by this definition. At a selection, the *eligible reward gap* Δ is the difference between the two largest rewards of the eligible records ($+\infty$ if at most one), and $\tau_t := \inf\{k : \Delta_k < t\}$ over the prediction steps that are selections. Under the fresh-embedding condition (no proposed embedding repeats an earlier embedding), the tree is the proposal graph of the rollout with the global visited convention; without it, the tree convention treats a repeated embedding as a new occurrence, which is the natural semantics for generated structures.

On a teacher trace (a prefix generated by the rule from the root token), the rule is the reward-ordered DFS teacher of the implicit tree: every occurrence is entered once, receives C candidate records, and is returned to at most once per valid child, so a trajectory has at most $(C+2)S$ tokens under the fresh-embedding condition and at most $(C+2)(B_{\mathrm{sel}}+1)$ under the budget of Definition E.3 below.

Definition E.3 (Search stopped at a selection budget)**.** Stop the teacher at its first accepting node or after B_{sel} fresh selections (returns and the initial root token are not counted), whichever is first; the accepting predicate reads decoded discrete registers only. The interface counts decoded fresh selections and excludes returns and the root. Termination also stops execution. A stopped prefix has at most $B := 2B_{\mathrm{sel}} + 1$ control steps (at most B_{sel} selections, as many returns, and one termination) and at most $T := (C+2)(B_{\mathrm{sel}}+1)$ columns.

Fix $t \in (0, 1]$ and use this context bound T. Set

$$0 \le \varepsilon < \tfrac{1}{2}, \qquad \pi(i) = e_i \in \mathbb{R}^T, \qquad \varepsilon_{\mathrm{pos}} = 0. \tag{E.1}$$

The Transformer is $F^{\mathrm{C}} = \psi_{\mathrm{out}} \circ \mathcal{A}^{\mathrm{c}} \circ \psi_{\mathrm{prep}}$, with $C+2$ heads in one attention stage. For $C = 2$, it has four heads.

Raw blocks and scratch. The raw layout is that of Definition E.1, of dimension $r_0 = 5d + 4T + 4C + 7$, including own-position metadata. Reserve $C + 1$ query blocks of dimension T, $C + 1$ preprocessed payload copies for read scratch, private register channels, and $O(C^2)$ route scalars. Thus $p = \Theta_C(d + T)$. Scratch is zero in all targets and every working scratch block is overwritten before use. Signed identity channels retain raw thoughts, embeddings and proposal values; private channels hold repaired registers. No repair is written into stored feedback.

Initial FFN and addresses. First apply $s_{\mathrm{bit}}(v) = \sigma(2v - 1/2) - \sigma(2v - 3/2)$ coordinatewise to $\phi, b, \nu, \iota, \mathsf{w}, n_{\mathrm{v}}, n_{\mathrm{c}}$ and $\pi_{\uparrow}, \pi_{\mathrm{fr}}, \pi_{\mathrm{fr}\uparrow}$ on every column. This repairs each intended bit within error $1/4$, in queries and value payloads. Preserve $z, z_{\mathrm{src}}, U, \mathsf{P}, V, r$, and clear h. All discrete-register symbols in the initial and output FFNs below refer to these private repaired values; embedding and thought symbols remain raw. Let $S_j e_i = e_{i+j}$ when $i + j \le T$, and zero otherwise. From the repaired counter one-hots compute $s = \mathbf{1}\{n_{\mathrm{c}} < n_{\mathrm{v}}\}$. This is the OR of the finitely many conjunctions $(n_{\mathrm{c}})_k \wedge (n_{\mathrm{v}})_l$ with $k < l$. Using bounded bit gates, form

$$u_j = S_j\big(s\pi_{\mathrm{fr}} + (1 - s)\pi_{\mathrm{fr}\uparrow}\big) \quad (1 \le j \le C), \qquad u_{C+1} = \pi_{\uparrow}.$$

The first C addresses name the current candidates at selection and the parent's candidates at return. The last names the linked anchor. Unused reads may be null, including all reads at exhausted root. Put $M = \log(T + 1)$ and compute internally

$$\ell = \mathrm{clip}_{[0,1]}\left(\sum_{k=1}^{T} \frac{\log(k+1)}{M}\pi_k\right), \qquad \Gamma_A(g, x) = \sigma(x - A(1 - g)) - \sigma(-x - A(1 - g)).$$

The logarithms are constants in a finite positional table. For each address write $(q_j)_i = \Gamma_1((u_j)_i, \ell)$ in query scratch. At position n, a valid $u_j = e_i$ gives $q_j = M^{-1}\log(n + 1)e_i$; a null address gives zero. The initial FFN keeps raw thoughts, embeddings and proposal values, repaired registers, and exact own positions, and zeros all read scratch.

Exact reads in one attention stage. The score matrix of head j has only the block MI_T from query q_j to source position. At a query with n visible columns and $u_j = e_i$, its logits are $\log(n + 1)$ at i and zero elsewhere. Let W project the preprocessed payload, $v_k = W\psi_{\mathrm{prep}}(x_k)$, and let A_i and U_0 be the addressed and uniform averages. Then

$$A_i = \frac{(n + 1)v_i + \sum_{k \ne i} v_k}{2n}, \qquad U_0 = \frac{1}{n}\sum_{k=1}^{n} v_k, \qquad 2A_i - U_0 = v_i.$$

This holds for arbitrary real payloads, also at $n = 1$. Let M_{unif} be the uniform head's score matrix. If L_j injects a payload into read block j, take

$$V_j = 2L_j W \quad (1 \le j \le C + 1), \qquad M_{\mathrm{unif}} = 0, \qquad V_0 = -\sum_{j=1}^{C+1} L_j W.$$

The uniform head supplies the same subtractand to every read block. The attention skip connection adds zero there, so the read block equals the target payload exactly. In the combined-matrix convention,

$$\max_h \|B_h\|_{\mathrm{op}} = \log(T + 1), \qquad \max_h \|V_h\|_{\mathrm{op}} \le \max\{2, \sqrt{C + 1}\}.$$

There is no read-tail parameter.

Reward and embedding comparisons. Write (z_j, V_j, r_j, ν_j) for the exactly read candidate fields and $B_\bullet$ for the linked-anchor fields. Eligibility is $e_j = \nu_j(1 - \mathsf{w}_j)$. For every $C \ge 2$, set

$$s_j = r_j - (2R + 1)(1 - e_j), \qquad g_{jk} = f_{0,t}(s_j - s_k), \qquad b_j = \sigma\left(\sum_{k \ne j} g_{jk} - (C - 2)\right).$$

On bounded legal inputs, $g_{jk}, b_j \in [0, 1]$. A unique eligible winner with top gap at least t has $b_j = 1$ and all other gates zero. Use the bounded-copy gadget Γ_A to dispatch

$$U_{\mathrm{out}} = q_{\mathrm{sgn}}\left(\sum_j \Gamma_{1/\sqrt{d}}(b_j, V_j)\right), \qquad z_{\mathrm{out}} = z_{\mathrm{src,out}} = \sum_j \Gamma_Z(b_j, z_j).$$

At bit-valued gates, these copy the chosen embedding and arbitrary bounded thought exactly. The protocol clips thoughts after prediction; the claim of exact execution stops at the first sub-threshold selection.

At return, the first C reads are parent records. Put

$$d_1(U,V)=\frac{1}{\sqrt{d}}\sum_k |U_k-V_k|, \qquad \widetilde{h}(U,V)=1-f_{\eta,1-\varepsilon-\eta}(d_1(U,V)), \qquad \eta=3(1-\varepsilon)/7.$$

Absolute values use $|v|=\sigma(v)+\sigma(-v)$. For canonical embeddings, $d_1(a_u,V)=1-\langle a_u,V\rangle$. If $\|U-a_u\|_2\le\eta$, then $|d_1(U,V)-d_1(a_u,V)|\le\|U-a_u\|_1/\sqrt{d}\le\|U-a_u\|_2$. The matching distance is at most η and every nonmatching canonical V has distance at least $1-\varepsilon-\eta$. The ramp is exact also at these endpoints. Its band width is $(1-\varepsilon)/7>1/14$ and its slope is $7/(1-\varepsilon)<14$. The matched slot is $h_j=\nu_j\widetilde{h}(U(x),V_j)$, gated by Γ_1. The same test $\widetilde{h}(\mathsf{P}(x),a_0)$ gives the root bit. Only one argument is perturbed; record V and a_0 are exact. No attention score performs slot lookup. These C comparisons use $O(Cd)$ units with norm-one reduction rows. Reward comparisons use exact interface rewards and contribute $O(C/t)$ to the Frobenius bound.

Output token fields. The three mutually exclusive routes are selection, nonroot exhaustion, and root exhaustion. They write:

(S) $z=z_{\mathrm{src}}=z_{\mathrm{out}}$, $\mathsf{P}=U(x)$, $U=U_{\mathrm{out}}$, $\phi=e_1$, $b=0$, $\iota=1$, $\mathsf{w}=0$, $n_{\mathrm{v}}=n_{\mathrm{c}}=e_0$, $\pi_\uparrow=\pi(x)$, $\pi_{\mathrm{fr}}=0$, $\pi_{\mathrm{fr}\uparrow}=\pi_{\mathrm{fr}}(x)$.

(R) $z=z_{\mathrm{src}}=B_{z_{\mathrm{src}}}$, $\mathsf{P}=B_{\mathsf{P}}$, $U=B_U$, $\phi=e_1$, $b=1$, $\iota=1$, $\mathsf{w}=B_{\mathsf{w}}\vee h$, $n_{\mathrm{v}}=B_{n_{\mathrm{v}}}$, $n_{\mathrm{c}}=\mathrm{shift}(B_{n_{\mathrm{c}}})$, $\pi_\uparrow=B_{\pi_\uparrow}$, $\pi_{\mathrm{fr}}=B_{\pi_{\mathrm{fr}}}$, $\pi_{\mathrm{fr}\uparrow}=B_{\pi_{\mathrm{fr}\uparrow}}$.

(T) The terminal token of Definition E.1: $\mathsf{P}=a_0$, $U=a_{\mathrm{eos}}$, $\phi=e_1$, $\iota=1$, $n_{\mathrm{v}}=n_{\mathrm{c}}=e_0$, all other predicted blocks zero.

Selection copies raw $U(x)$ into P; return copies the linked raw $U,\mathsf{P},z_{\mathrm{src}}$ and the repaired mask, counts and pointers. Raw embedding-copy gates use coordinate bound $1+\eta$, since each admitted embedding has norm at most $1+\eta$. The fresh V sign snap acts only on an already canonical selected proposal; it is never applied to raw U,P copies. Routes (S) and (R) set V,r,ν to zero; every route zeros scratch. These are executor outputs, not interface resets. The interface appends the next own-position as metadata. It also performs the discrete acceptance and budget tests. All canonical constant blocks, including route flags, are guarded before routing by $\kappa=s_{\mathrm{bit}}(\iota_{\mathrm{raw}})$, using $\Gamma_{|c_i|}(\kappa,c_i)$ on each nonzero coordinate of the constant c. Every internal bit and route is clamped to $[0,1]$ before use; these clamps and guards fix all legal-input computations.

Depth, sparse support, and norms. We use the resource bounds of Lemma A.6, keeping each reduction, scalar ramp, and broadcast in its own stage. The initial FFN first repairs the listed registers coordinatewise, preserving the raw continuous and embedding blocks. It then computes counter conjunctions, their OR, and the clipped positional-table scalar in parallel; it then gates the two pointer choices, adds and shifts them, and writes the query gates. Preprocessed payloads and exact positions are retained by signed identity channels; read scratch is zero. These are $O_C(1)$ successive gadget stages.

Output first forms clamped eligibility bits and absolute embedding differences in parallel; it next reduces embedding differences, computes equality and reward ramps, and forms slot and route gates. Subsequent stages compute the C-slot dispatcher, gate thought and embedding coordinates, round the embedding, update masks and counters, guard constants, dispatch the three routes, and clear scratch. The dependencies just listed have $O_C(1)$ depth. Absolute values are $\sigma(v)+\sigma(-v)$, scalar clips are differences of two ReLUs plus a constant, and all other nonlinear operations are the explicit gadgets (i)–(vii) of Lemma A.6, or the displayed Γ identities. Signed copies preserve every needed raw or intermediate block between stages. There are $O_C(1)$ such vector blocks and $O_C(1)$ scalar registers, so the number of retained coordinates and the hidden width are $O_C(p)$; no numerical depth or hidden-width multiplier is needed.

Each coordinate operation or copy has bounded fan-in; each reduction and broadcast has $O(p)$ entries, and each shift is a partial permutation. Thus every affine matrix and bias has $O_C(p)$ nonzero entries. Their magnitudes are $O_{C,R,Z}(1)$ except for the $O_C(1)$ scalar coefficients of the reward

ramps, which are $O_C(t^{-1})$. In particular, the positional-table coefficients $\log(k+1)/\log(T+1)$ are at most one and the embedding-distance coefficients are $d^{-1/2}$. Repair slopes are two and threshold slopes are strictly less than fourteen. Neither reduction is composed affinely with a broadcast. These resource bounds therefore give

$$\max_i\{\|W_i\|_{\mathrm{F}}, \|b_i\|_2\} = O_{C,R,Z}(\sqrt{p}+t^{-1}), \qquad \Lambda_{\mathrm{ffn}} = O_{C,R,Z}(\sqrt{p}+t^{-1}).$$

This is a sparse witness in the containing norm-bounded causal class. Its fixed sparse support includes the repair and threshold stages. The exact attention maps still have $\Lambda_B = \log(T+1)$ and $\Lambda_V \le \max\{2, \sqrt{C+1}\}$.

For completeness, the linear growth conditions of Lemma A.6 hold on arbitrary real inputs. Coordinate repair obeys $|s_{\mathrm{bit}}(v)| \le 2|v|$. The initial FFN's shifted, gated addresses have norm bounded by a constant times the raw pointer norms, and $|\Gamma_1(u_i, \ell)| \le 2|u_i|$ prevents the positional reduction, whose norm can grow with T, from being emitted. In output, $\|\Gamma_Z(b_j, v)\|_2 \le 2\|v\|_2$ for each clamped gate, so the C-way dispatch has a linear growth coefficient depending only on C. Embedding dispatch and rounding have the same linear growth bound. The matched-slot gate obeys $|\Gamma_1(\mathrm{clip}_{[0,1]}(\nu_j), h)| \le 2|\nu_j|$ for $h \in [0,1]$; the mask OR on clamped inputs is bounded by their sum. Counter shifts and all remaining nonconstant writes are copies, partial permutations, or gated sums of raw or read blocks. The only other writes are bounded-norm constants guarded by κ, with $|\kappa| \le 2|\iota_{\mathrm{raw}}|$. A fixed number of blocks and stages thus gives $\|\psi(x)\|_2 \le O_{C,R,Z}(1)\|x\|_2$ for both complete maps, independently of p, t. Their zero-input values are zero. Together with the value-map bound this keeps the executor's bounded-column output envelope $O_{C,R,Z}(1)$.

Lemma E.4 (Margin-exact candidate selection)**.** *On any canonical anchor satisfying the link, mask and counter invariants, or an admitted raw anchor whose repaired registers satisfy them, a nonempty eligible set with top gap at least t is dispatched exactly, including its embedding and both thought blocks, for every $C \ge 2$.*

Proof. Exact fixed-offset reads supply every candidate record. Ineligible adjusted scores are at most $-R-1$, while eligible scores are at least $-R$. Since $t \le 1$, the winning eligible slot has $g_{jk} = 1$ against every other slot, even for singleton selection. Any losing slot has a zero comparison against the winner, so its sum is at most $C-2$. Thus the b_j are exactly the winner's one-hot. Bit-gated copies give the embedding and both arbitrary bounded thought blocks. □

Lemma E.5 (Exact addressed read)**.** *At a query with $n \le T$ visible columns, let a used pointer be $u = e_i$ for $1 \le i \le n$. The addressed head with target logit $\log(n+1)$, together with the common uniform head, recovers every preprocessed payload coordinate of column i exactly. This applies to candidate offsets and linked anchors, including raw continuous thoughts and embeddings and privately repaired positional content pointers.*

Proof. The own-position table gives $q = M^{-1}\log(n+1)e_i$. Orthonormal source positions give zero logits off i, so the softmax denominator is $(n+1)+(n-1) = 2n$. For arbitrary payloads $v_1, \dots, v_n$,

$$2\frac{(n+1)v_i + \sum_{k\ne i} v_k}{2n} - \frac{\sum_{k=1}^n v_k}{n} = v_i.$$

Both terms use the same preprocessed-payload projection W. The initial FFN zeros the read scratch, so the residual contributes nothing, and the signed value maps implement this identity in each scratch block. The fresh-ID and backlink invariants make every used address a visible column. Null pointers are never used by the active route. No positional write-back or limiting temperature is needed; content pointers are repaired privately before query formation. □

Lemma E.6 (Exact return-slot comparison)**.** *Let a well-formed canonical or admitted raw exhausted nonroot anchor have $\|U(x) - a_u\|_2 \le \eta = 3(1-\varepsilon)/7$, where a_u is exactly one valid candidate embedding in its parent visit. The exact parent-record reads and threshold comparisons give precisely that slot's one-hot. The repaired mask OR and completed-counter shift are exact.*

Proof. Repaired parent pointers and fixed offsets name the C parent records. Lemma E.5 recovers their canonical candidate embeddings. The displayed d_1 perturbation bound gives equality bit one at a_u and zero at all distinct dictionary embeddings. Validity gates remove invalid slots. The linked

anchor's repaired mask records the state before this child's selection, so the slot has not yet been marked. OR adds it once and the repaired one-hot shift increments the completed count. □

Scope constants. Let R project onto all predicted coordinates except $z, z_{\mathrm{src}}, U, \mathsf{P}$, including inactive V, r, h. It is a coordinate group, not a feedback map. Fix

$$a = 1/4, \qquad \eta = 3(1-\varepsilon)/7, \qquad L_{\mathrm{copy}} = 2(C+1), \qquad \rho'' = \min\{a, \eta/L_{\mathrm{copy}}\}.$$

The admissibility bounds limit joint R error by a and each of the two embedding errors by η, relative to the canonical rule decision on the decoded prefix. The bounds are strict.

Definition E.7 (Copy labels and recursive scope)**.** First define a total admissibility indicator $A(E) \in \{0,1\}$ by finite replay of the supplied prefix. Set $A(\dagger) = 0$ and reject malformed or nonfinite inputs. Start with the exact root and its prescribed records if execution continues. At each stage:

1. Before admitting an anchor, check that its decoded prefix is well-formed and has not stopped: phases, links, masks, counts, visible addresses, proposal bounds and record construction satisfy Definitions E.1–E.3, and both actual thoughts have norm at most Z. Check route prerequisites without evaluating H_{C}: a due selection has an eligible winner and gap at least t; a nonroot exhaustion has a visible linked parent anchor and parent records with a unique valid matching slot. Exhausted root needs no linked parent. Reject if a check fails.
2. If the prefix contains the next emission $\widehat{o}_i$, compute $o_i^{\mathrm{rule}} = \mathsf{H}_{\mathrm{C}}(S(E_{i-1}))$ only from this already admitted shorter prefix. Require the decoded equality of Definition E.2, the thought bounds, and
$$\|\mathsf{R}(\widehat{o}_i - o_i^{\mathrm{rule}})\|_2 < a,$$
$$\|U(\widehat{o}_i) - U(o_i^{\mathrm{rule}})\|_2 < \eta, \qquad \|\mathsf{P}(\widehat{o}_i) - \mathsf{P}(o_i^{\mathrm{rule}})\|_2 < \eta.$$
Reject on failure before acceptance, termination, budget stopping, action dispatch, record insertion or any front-end call from this emission. Otherwise, process its decoded stop and action. If execution continues after a fresh selection, verify the next C records, including raw copy fields, exact metadata and proposals from the actual clipped thought and decoded address. No continuation is permitted after a normal stop.
3. Continue the replay. Set $A(E) = 1$ exactly when it ends at an admitted anchor before stopping, including the currently due gap check; otherwise set $A(E) = 0$.

All historical rule evaluations use strictly shorter admitted prefixes. The check depends only on the prefix and fixed interface, not on a clean teacher trajectory or the candidate Transformer being fitted. Labels and scope are evaluable with oracle access to g_{pre} and exact real comparisons; no recursive computability is asserted.

Only when $A(E) = 1$, define the copy label $o_{\mathrm{cp}}(E)$. Let x be the last column, and apply H_{C} to $S(E)$ to determine the route and selected slot $i^\star$. On the return route only, let y be the linked parent anchor identified by the validated pointer. The label uses the rule's canonical R coordinates and actual record/source thoughts. Its embeddings are

	U	P
selection	$V_{i^\star}$	$U(x)$
return	$U(y)$	$\mathsf{P}(y)$
termination	a_{eos}	a_0.

At selection, both thoughts equal the selected record's thought; at return, both equal the linked anchor's raw source $z_{\mathrm{src}}(y)$. Masks, counts and pointers are canonical repaired values, not raw copies. The label is undefined on rejected prefixes. Define the scoped loss by

$$\ell_{\mathrm{cp}}(F; E) = \begin{cases} \frac{1}{2}\|F(E)_{:,-1} - o_{\mathrm{cp}}(E)\|_2^2, & A(E) = 1, \\ 0, & A(E) = 0. \end{cases}$$

The zero branch evaluates neither the Transformer nor the partial label. Own-position metadata is excluded from prediction and loss.

Executor setting. Assume equation E.1. For $\|z\|_2 \le Z$, let the fixed front end return thoughts of norm at most Z, rewards of magnitude at most R, and pairwise distinct valid child embeddings different from a_0 and a_{eos}, in any subset of the C slots. Fix $t \in (0, 1]$, fixed branching C, and the stopped problem with budget B_{sel} with $T = (C+2)(B_{\mathrm{sel}}+1)$.

Theorem E.8 (Causal one-layer thought executor; closed-loop correctness)**.** *There is $F^{\mathrm{C}} \in \mathcal{F}^{\mathrm{c}}_{\mathrm{TI}}(\mathbf{\Lambda})$ with one attention stage, $C+2$ heads, token width $p = \Theta_C(d+T)$, FFN depths $O_C(1)$, FFN hidden width $O_C(p)$, and norm bounds*

$$\Lambda_B = \log(T+1), \qquad \Lambda_{\mathrm{ffn}} = O_{C,R,Z}(\sqrt{p} + t^{-1}), \qquad \Lambda_V, c_{\mathrm{ffn}} = O_{C,R,Z}(1),$$

such that:

(a) (Canonical prefixes.) *On every well-formed canonical anchor with an eligible reward gap at least t if selection is due, the output equals H_{C} in every predicted coordinate, including both thought blocks and inactive zeros. Returns and termination require no reward-gap condition.*

(b) (Closure.) *On every instance, the rollout from the root agrees exactly with the rule until acceptance, τ_t, termination, or budget. Thoughts stay in the radius-Z ball. Before this stop, each selection copies its selected record's thought exactly; each return restores the fresh occurrence's thought exactly. The Transformer is causal and KV-cache-consistent.*

(c) (Admitted raw prefixes.) *For every prefix admitted by Definition E.7, $F^{\mathrm{C}}(E)_{:,-1} = o_{\mathrm{cp}}(E)$ and $\ell_{\mathrm{cp}}(F^{\mathrm{C}}; E) = 0$, before clipping in every predicted coordinate.*

Proof. Resources and causality. The preceding construction specifies the private repairs, simultaneous reads, threshold comparisons, and output FFN. On canonical prefixes, repairs preserve the registers. Its composition argument gives the claimed bounds, including linear growth on arbitrary real inputs. The fixed-offset and linked-anchor addresses needed by each route are already present before attention. All sources are visible, and the FFNs are columnwise, so Lemma A.3 gives KV-cache consistency.

Invariants. Induct over prediction steps with: (I1) the prefix is well-formed; (I2) every thought has norm at most Z; (I3) each backlink names the parent anchor, each fresh identifier names its fresh token, the parent fresh identifier is correct, and each mask marks exactly the slots whose children have returned; (I4) all earlier control outputs agree exactly with H_{C}. The root and its interface-written records satisfy these invariants. For every occurrence, the counters give the valid and completed slot counts, since each return marks a distinct eligible slot.

Route and selection. The counter comparison therefore selects exactly when an eligible record exists. The root comparison is exact because no valid child has embedding a_0. At selection, the current fresh identifier and offsets name the current records. Lemma E.5 makes their entire payloads exact, and Lemma E.4 makes the winning slot's gate a bit when the gap is at least t. The output FFN copies its thought into both thought blocks and its embedding into U, copies $U(x)$ into P, and installs the backlink, parent fresh identifier, zero mask, and initial counters exactly. The copied thought has norm at most Z. The interface writes the selected occurrence's records from this thought with its new fresh ID. Thus (I1)–(I4) persist.

Return. At nonroot exhaustion, the linked anchor is an earlier parent anchor. Its raw payload is exact by Lemma E.5. The returning child's embedding is the embedding actually chosen from a unique valid parent slot. By Lemma E.6, the output FFN marks precisely that slot, shifts the completed count, and copies the parent's other registers. Both output thoughts equal that anchor's source register. Initially the source register is the occurrence's fresh thought; every previous return restored it exactly. It therefore remains the fresh thought, with norm at most Z. The induction is restored at the parent.

Termination. Exhausted root takes the remaining route and emits the fixed terminal token. Raw feedback and thought clipping fix these exact emissions. Every executor route clears scratch and all inactive fields; own-position metadata is supplied by the interface. This proves (b). The same local argument applies to every well-formed canonical prefix (the return rule copies the linked source register), and proves (a) on canonical prefixes, including teacher returns after earlier small gaps: the links, masks, and counters still satisfy the invariants.

Admitted raw prefixes. The joint R bound gives coordinate error below $1/4$ for $\phi, b, \nu, \iota, \mathsf{w}, n_{\mathrm{v}}, n_{\mathrm{c}}$ and all three pointers. Copies into records preserve these errors. The initial FFN repairs them before queries and value reads, preserves raw thoughts, embeddings, V, r, and clears scratch. Lemma E.5 therefore reads exactly the intended preprocessed columns. Records have exact proposal embeddings, validity and rewards, and scope supplies the current gap. Lemma E.4 copies the winning record's canonical embedding and actual thought, with raw $U(x)$ as the new parent embedding. At

return, $\eta = 3(1-\varepsilon)/7$ and Lemma E.6 give the exact slot and root bits. The output FFN copies linked raw embeddings and source thought, but canonical repaired masks, counters and pointers. Termination and inactive outputs are canonical. These are precisely the copy labels. The decoded address passed to the front end lies in its original domain, so Assumption 1 supplies every proposal bound used here. No continuity or quantization of thoughts is needed. □

E.2 Scope of stopped execution

Theorem E.8 certifies agreement with the search rule until acceptance, τ_t, termination or budget, for every $C \geq 2$. This concerns the structure generated by the emitted thoughts. Reaching an external goal additionally requires that the front end generate candidates that include it. Closed-loop correctness allows search orders absent from the teacher while requiring correct procedure on the rollout's own tree; path success instead requires agreement with that teacher.

Remark E.9 (Teacher-path departure and closed-loop correctness). For arbitrary base Transformers, Theorem F.8 proves departure from the teacher's decoded trace. It does not require their thoughts or generated proposals to equal the teacher's. If no discrete departure has occurred, the two raw fields copied into each fresh anchor still give the recurrence for $(z_{\mathrm{src}} - U)_1$ used in the counterexample. For the exact executor, Corollary F.9 additionally proves canonical proposals and closed-loop correctness through the first $H-1$ selections on its high-gap event. At the next selection, the root replaces an eligible child, so on-policy labels detect the error. No claim about reaching the goal after departure is needed.

E.3 On-policy learning

Under raw feedback, the rule is determined by the learner's prefix and the front end. The training target retains raw embedding copies where the executor copies them. We establish the common error-to-execution lemma, then prove Theorem 3 for one shared Transformer.

E.3.1 Common prediction domain and execution bound

Prediction domain. Fix confidence $\delta_0 \in (0,1)$ and a norm-bounded causal class containing the robust F^{C} at threshold t, at any fixed attention depth. A fixed sparse template must include the construction's repair and threshold support. On the scoped numerical domain,

$$B_x = O_C(1 + R + Z + a + \eta), \qquad B_y = O_C(1 + Z + \eta).$$

Indeed, embeddings have norm at most $1+\eta$, the joint register error is below a, and canonical columns have only $O_C(1)$ nonzero one-hot and bit blocks. Raw record copies preserve these bounds. Use the covering-number bound under d_∞ on this domain, $\log \mathcal{N}(\alpha) \leq D \log(1 + A/\alpha)$, $D \geq 1$, for example $A = A_{\mathrm{pred}}$ from Lemma C.3. Here, D and A come from the bound for the dense or fixed sparse class on that scoped domain. Lemma C.3(ii) applies in particular at the scale $\alpha = \rho''/3$ used below. These bounds apply to the aggregated on-policy protocol below. One may embed the one-stage executor in a deeper class using zero value matrices and identity FFNs in the extra stages, provided the architecture and norm bounds include them: width at least $2p$, layer radius at least $\sqrt{2}$ and growth coefficient at least one suffice for signed ReLU identities.

Lemma E.10 (An exit forces a copy error)**.** *For any sequence of Transformers run with the admission and emission checks of Definition E.7, pad stopped rollouts by †, and let $I_k = A(E_{k-1})$ and*

$$e_k = \begin{cases} F_k(E_{k-1})_{:,-1} - o_{\mathrm{cp}}(E_{k-1}), & I_k = 1, \\ 0, & I_k = 0. \end{cases}$$

Let $\mathcal{E}$ denote a violation of closed-loop correctness at a step $k \leq B$ before stopping, $k < \tau_t$, or an exit from the scope tolerances before that stop. Then, deterministically,

$$\mathcal{E} \subseteq \bigcup_{k=1}^{B} \{I_k = 1,\ \|e_k\|_2 \geq \rho''\}.$$

This also applies when the same Transformer is used at every step.

Proof. Embedding-copy paths. An occurrence v begins at a fresh selection, even if its embedding appeared earlier. Let a_v be its canonical entry embedding, and $\xi_{k,U}, \xi_{k,P}$ the indicated components of e_k. Before the first failed emission check, each return marks a distinct slot, so an occurrence has at most C returns. Its current embedding at its r-th return satisfies

$$U_{v,r} - a_v = \sum_{j \in I(v,r)} \xi_{j,U}, \qquad |I(v,r)| \le C + 1.$$

Indeed, its fresh selection copies a canonical record embedding and adds one new copy error; each return copies the linked occurrence's previous U and adds one new copy error. The root starts with zero error. For nonroot v, let r_v count returns to its parent before v is selected. The initial parent embedding copies that parent's current U; subsequent returns copy v's previous P. Hence

$$\mathsf{P}_{v,r} - a_{\mathrm{parent}(v)} = \sum_{j \in I(\mathrm{parent}(v), r_v)} \xi_{j,U} + \sum_{j \in J(v,r)} \xi_{j,P}, \qquad |I| + |J| \le 2(C+1).$$

Root parent embeddings have no inherited error. The new copy errors are from admitted input steps; interface record copies add none. Fresh U and both terminal embeddings have only the current copy error. The identities remain valid at the first failed emission, since its input and copy label are still admitted. Copying P on a return does not pass it down another level of the tree.

First exit. A first joint register exit has $\|\mathsf{R} e_k\|_2 \ge a$, since its label registers are canonical. For a first embedding exit, the triangle inequality on the corresponding path of at most L_{copy} new copy errors gives

$$\eta \le \Big\| \sum_j \xi_j \Big\|_2 \le L_{\mathrm{copy}} \max_j \|\xi_j\|_2.$$

Thus some admitted step, possibly earlier than the exit, has $\|e_j\|_2 \ge \min\{a, \eta / L_{\mathrm{copy}}\} = \rho''$. Clipping thoughts changes no register or embedding error.

Before any numerical exit, every register and inactive nullable V is within distance $1/2$ of its prescribed value, and both embeddings are in their intended cells because $\eta < 1/2 < \rho_{\mathrm{code}}$. The decoded emission is therefore the rule's decision. The interface uses its decoded address and actual clipped thought, so the bounded front end supplies valid new records satisfying the prefix invariants. Inactive r, h remain within the monitored joint register bound. Induction from the exact root shows that an incorrect decision cannot precede a numerical exit. Emission checks precede acceptance, termination and budget stopping, so an incorrect final emission is also charged. Stopping before a due low-gap selection evaluates no label there. These observations prove the event inclusion. □

E.3.2 Shared-model on-policy imitation learning

Definition E.11 (Aggregated on-policy imitation learning)**.** Use Setting 3, with the admission indicator $A(E)$ and copy label $y(E) := o_{\mathrm{cp}}(E)$ of Definition E.7. Set $\eta := 3(1 - \varepsilon_{\mathrm{emb}})/7$ and $\rho'' := \min\{1/4, \eta/[2(C+1)]\}$, using the same coordinate-block error bounds as in that definition. All numerical bounds, the threshold t, the horizon B, the model class, and $N \ge 1$ are fixed before sampling. Instances have a fixed law P_X on a standard Borel space. The front end is a fixed jointly Borel function of the instance and its numerical inputs, with no state carried between instances or rounds; all finite decoding ties use fixed rules. These conventions require no continuity or Lipschitz bound on the front end. The test instance is independent of the entire training history.

The class has a fixed compact feasible parameter set Θ with all declared coefficients trainable. Fix a measurable squared-loss ERM selection rule, whose existence is proved below. Start with $\mathcal{D}_0 = \varnothing$ and an arbitrary fixed $F^{(1)} \in \mathcal{F}$; the selection rule may be set to return this model on empty data.

At round r, draw N i.i.d. instances from P_X, independently of all previous data, and use the same $F^{(r)}$ at every prediction step. Run raw feedback from the exact root with the admission monitor of Definition E.7, up to B steps and the prescribed stops. Collect every admitted input and its label, including the input to the first failed emission check, as a multiset $\mathcal{S}_r$. Evaluate the prediction before processing any acceptance, termination, budget stop or scope exit it triggers. No off-scope label or continuation is evaluated, and a currently due sub-threshold selection is not labeled. For each collected pair, test $\|F^{(r)}(E)_{:,-1} - y(E)\|_2 < \rho''$ on all predicted coordinates. If all tests

pass, return $F^{(r)}$ unchanged, without fitting to this batch. Otherwise, set $\mathcal{D}_r = \mathcal{D}_{r-1} \cup \mathcal{S}_r$ with multiplicity and choose $F^{(r+1)}$ by the fixed ERM rule for

$$\widehat{\mathcal{R}}_r(F) = \frac{1}{2\max\{1, |\mathcal{D}_r|\}} \sum_{(E,y)\in\mathcal{D}_r} \|F(E)_{:,-1} - y\|_2^2.$$

Inputs and targets are fixed throughout each minimization. The monitor and labels are used only for training and evaluation, not to correct deployment outputs. Empty data have zero loss; no initial search competence is assumed.

Measurability and ERM selection. For each fixed input length, $(\theta, E) \mapsto F_\theta(E)$ is jointly continuous: matrix operations, fixed-mask softmax and ReLU are continuous, and the network is their finite composition. A fixed sparse support and norm constraints are closed; each FFN linear growth constraint is also closed for every fixed input, hence their intersection is closed. Finite-dimensional norm bounds make Θ compact, and executor membership makes it nonempty. For finite data d, the empirical loss $L(d, \theta)$ is jointly continuous on each fixed-size data component. The front end and discrete decoding are not reevaluated during this optimization.

A measurable minimizer can be constructed explicitly. Enclose Θ in a fixed closed cube and repeatedly subdivide it into closed dyadic cubes with a fixed ordering. For each cube Q, the minimum $v_Q(d) = \min_{\theta\in\Theta\cap Q} L(d, \theta)$ is continuous on each data component when the intersection is nonempty; otherwise set $v_Q = +\infty$. At each subdivision, choose the first child whose minimum equals the global minimum. The chosen parent contains a global minimizer and its children cover it, so such a child always exists. This choice is measurable. The sets of global minimizers in the chosen cubes are nonempty nested compact sets with diameters tending to zero. Their common point minimizes the loss and is measurable as the limit of the chosen cube centers. On the empty-data component, any fixed feasible model may be selected.

Finite data are encoded in the countable disjoint union of the Euclidean spaces for their sizes and token lengths. Finite replay, comparisons and the jointly Borel front end make admission, labels and monitored rollouts jointly measurable in model parameters and instances. Enumerating pairs by round, instance and step therefore gives measurable datasets and ERM iterates. Integration over the fixed law P_X makes each model's rollout error probability measurable as well. This establishes the conditional probabilities used below; it does not assert an efficient ERM solver.

Theorem E.12 (Finite-round learning of a shared Transformer)**.** *Under Definition E.11, let Ω be the fixed set of all admitted prefixes and set*

$$d_\infty(F, G) := \sup_{E\in\Omega} \|F(E)_{:,-1} - G(E)_{:,-1}\|_2, \qquad K := \mathcal{N}(\rho''/3, \mathcal{F}, d_\infty) \le (1 + 3A/\rho'')^D.$$

If the admitted domain is empty, set $d_\infty = 0$ and $K = 1$; the first batch then passes and the scoped failure event is empty. Every prescribed ERM exists. There are at most $K - 1$ failed rounds, and the protocol returns a tested Transformer $\widehat{F}$ by round K, using at most NK instances and BNK labels. With probability at least $1 - \delta_0$ over the instances sampled during training, its probability on an independent test instance of a closed-loop violation at a step $k \le B$, $k < \tau_t$, or a scope exit before acceptance, termination, the budget or a due sub-threshold selection, is at most

$$1 \wedge \frac{\log K + \log(1/\delta_0)}{N} \le 1 \wedge \frac{D\log(1 + 3A/\rho'') + \log(1/\delta_0)}{N}.$$

Test probabilities are conditional on the realized training history. The returned Transformer is used at every deployment step without a teacher or write-back. This proves Theorem 3.

Proof. Interpolation and finite termination. Theorem E.8(c) gives $F^{\mathrm{C}}(E)_{:,-1} = y(E)$ on every admitted prefix, regardless of which model generated it. Hence every accumulated dataset is realizable, and each ERM has zero loss on every stored pair. Compactness and continuity give existence, with the measurable selection rule specified above. For each failed round r, choose a witness $E_r \in \mathcal{S}_r$ with $\|F^{(r)}(E_r)_{:,-1} - y(E_r)\|_2 \ge \rho''$. Every later model and F^{C} fit this witness exactly. Thus the failed models together with F^{C} are pairwise separated by at least ρ'' in d_∞. A ball of radius $\rho''/3$ has diameter at most $2\rho''/3$ and contains at most one of them. A K-ball cover therefore permits at most $K - 1$ failed rounds. The next round must pass, proving termination and the instance and label counts. The cover is on the fixed admitted domain, not on a model-dependent input distribution.

Certification at an adaptive stopping round. Condition on the history before round r, so $F^{(r)}$ is fixed. Let q_r be the probability that its monitored rollout on a new instance contains an admitted prediction error at least ρ''. The new batch is independent of that history, so its conditional pass probability is $(1-q_r)^N \le e^{-Nq_r}$. Put $s = \log K + \log(1/\delta_0)$. At any reached round, the conditional probability of passing while $q_r > s/N$ is at most e^{-s}. Taking expectations and a union bound over at most K rounds bounds the probability of ever accepting such a model by $Ke^{-s} = \delta_0$. This includes the data-dependent stopping round r_*. We do not assume that the selected $\widehat{F}$ is independent of the passing batch, nor that steps within an instance are independent.

Execution. Lemma E.10 applies to the constant sequence $(F^{(r)}, \dots, F^{(r)})$. A correctness violation or scope exit implies an error at least ρ'' at some admitted step at or before that event, not necessarily at the exit step itself. The monitored and ordinary rollouts agree up to their first violation or prescribed stop, and the input to the offending emission is evaluated before processing its stop. The error witness therefore belongs to their common prefix. Consequently, deployment failure has probability at most q_r. The protocol returns the tested $F^{(r_*)}$ without updating it, so the certification bound applies. Capping by one and using the covering-number bound proves the stated inequalities. No front-end Lipschitz condition, thought quantization or claim about SGD is used. □

Scope of the on-policy guarantee. The admission conditions concern registers and copied embeddings, not closeness to teacher thoughts. The front end requires neither thought quantization nor a Lipschitz bound. The constants D, A, ρ'' retain their dependence on C. The round bound is an ERM-oracle upper bound, not an optimization guarantee or a lower bound on the resources needed.

F TEACHER FORCING UNDER THE QUANTIZED FRONT END

F.1 THE FINITE-COVER SUFFICIENT CONDITION

Use the quantized front end of Section 5.2 on $\mathcal{E}_H^{(C)}(\bar{\zeta})$ and the original thought interface, with $B_{\mathrm{sel}} \ge 1$, $B = 2B_{\mathrm{sel}}+1$, and $T = (C+2)(B_{\mathrm{sel}}+1)$. Fix any predeclared constant attention depth, constant FFN depths, and a norm-bounded class with finite covering numbers under d_∞ on all bounded numerical prefixes. The main text writes $A = A_{\mathrm{pred}}$. Only the covering-number bound and the input bound of Lemma C.3 are used; $D, A_{\mathrm{pred}}, L_{\mathrm{in}}$ absorb depth. The modification W_a below adds only FFN stages and preserves the base Transformer's attention depth.

Input stability from the parameter-cover proof. Use the covering-number constant

$$A_{\mathrm{pred}} := 4\sqrt{\max\{p, W_{\mathrm{ffn}}\}}\,\Lambda_{\max} A_F$$

of Lemma C.3 for the dense matrix parameterization, whose parameters include every B_h; for the sparse template write A_{pred} for the enlarged constant of Lemma C.3(iii), which carries the source radius and the parameter-map Lipschitz constant. Declare the column radius on numerical prefixes, including off-canonical discrete blocks and inactive V, r, h, by

$$B_x \ge B_{\mathrm{teach}} + \rho/2, \qquad B_{\mathrm{teach}} := \sup_{\mathcal{T},k,i} \|(E_k(\mathcal{T}))_{:,i}\|_2.$$

The supremum includes records and the original own positions. It is $O_C(1+R+Z)$. Both the input Lipschitz bound and the covering-number bound under d_∞ hold on this ball, including causal prefixes. Prediction errors are measured on the predicted coordinates included in the loss, including all predicted blocks and scratch, excluding supplied own positions; coordinate projection preserves both bounds. For one attention stage with identical parameters, the attention and FFN input bounds compose to give, for aligned prefixes,

$$\|F(E)_{:,-1} - F(E')_{:,-1}\|_2 \le L_{\mathrm{in}} \|E - E'\|_{2,\infty}, \qquad L_{\mathrm{in}} := \Lambda_{\mathrm{ffn}}^{L_{\mathrm{pre}}+L_{\mathrm{out}}} \big[1 + J\Lambda_V(1 + 2\Lambda_B B_\circ^2)\big],$$

where $L_{\mathrm{pre}}, L_{\mathrm{out}}$ are the initial and output FFN affine depths (the modification's additive scale a is unrelated). For $q = L_{\mathrm{att}}$ attention stages, the same composition gives

$$L_{\mathrm{in}} := \Lambda_{\mathrm{ffn}}^{\sum_{r=0}^{q} L_r} \big[1 + J\Lambda_V(1 + 2\Lambda_B B_\circ^2)\big]^q,$$

where L_r is the affine depth of FFN r, and $B_\circ$ uses this depth. This includes both thought blocks of the last query column as well as all earlier columns. It also implies a bound with the sum of column

perturbations and the same constant. Explicitly, score differences are at most $2\Lambda_B B_\circ \|E - E'\|_{2,\infty}$ at an attention input; the softmax Lipschitz constant one, value transport and residual give the bracketed factor. The FFNs contribute the products of their per-layer operator norms. No Lipschitz property of the front end is used in this input estimate. At $B_{\mathrm{sel}} = O_C(H)$, under the polynomial norm bounds for the exact executor and W_a below, fixed depths and fixed $a > 0$,

$$\log(1 + A_{\mathrm{pred}}) + \log(1 + L_{\mathrm{in}}) = O_C(\log(NH/\delta)).$$

The declared class includes the polynomial FFN allowance for W_a. For Definition A.4, use $p' = p + 2$ in these architecture constants; the reserved coordinates are zero at inputs, targets and emissions. The cover and input bounds are inherited from that same ambient architecture, as proved in Lemma F.7. A fixed sparse TI support accommodating these operations has $D = O_C(H)$ and its enlarged A_{pred} is still polynomial in NH/δ. For varying radii or depths, use the displayed constants, not this specialized order. Use ρ for the quantization radius of Section 5.2. The embedding-distance identity in Lemma C.4 and incoherence $\varepsilon < 1/2$ give $\sqrt{(1-\varepsilon)/2} \le \rho \le \sqrt{(1+\varepsilon)/2}$. Thus $\rho/2 < \sqrt{3}/4 < 1/2$, strictly below every discrete decoding radius. The rollout induction below uses this discrete decoding radius and the lookup ball of radius ρ separately.

Trace margin and pre-sample norm bounds. Let $m(X)$ be the least eligible reward gap at a queried selection of the stopped teacher, with singleton and empty comparisons assigned $+\infty$. Then

$$\mathbb{P}\{m(X) < t\} \le \min\{1, \kappa_C B_{\mathrm{sel}} t\}.$$

Indeed, at each visited internal node the remaining eligible children form a suffix of its reward order. Every nonsingleton selection therefore uses an adjacent gap. The event of any small gap is bounded by the sum over first-entry indicators times the event that any adjacent gap is small. To justify the first-entry bound despite goal stopping, condition on all preferred children, which fixes the goal, and reveal an outgoing Gaussian block only at its node's first query. Before that query, the DFS decisions and stopping use only previously revealed blocks and the fixed goal. Each newly queried block is therefore conditionally fresh. Lemma B.3(iii) bounds its small-adjacent-gap event by $\kappa_C t$, uniformly in its preferred child. Including the root, there are at most B_{sel} queried first entries: a fresh selection that exhausts the budget receives no records or query. Summing these conditional bounds, with zero padding after stopping, and then removing the conditioning proves the display. This covers later selections without conditioning on their occurrence or multiplying by $C - 1$. The same argument also bounds the expected number $N_t(X)$ of sub-threshold selections. Each adjacent gap is used at most once at its node, so charge $N_t(X)$ to the sum of small-adjacent-gap counts at first queries. Lemma B.3(iii) bounds each conditional expected count by $\kappa_C t$, giving $\mathbb{E}N_t(X) \le \kappa_C B_{\mathrm{sel}} t$.

For $N \ge 1$ and $\delta \in (0, 1)$, declare before sampling

$$t_N = \frac{\delta}{2\kappa_C N B_{\mathrm{sel}}} = \frac{\delta}{C(C-1)NB_{\mathrm{sel}}}.$$

The chance that any of the N training traces has margin below t_N is at most $\delta/2$. Choose compact norm balls containing the exact executor at this threshold and impose the closed linear growth constraint. The architecture is fixed; norm radii are declared as functions of N, δ before sampling. At $B_{\mathrm{sel}} = 5H$, sufficient norm bounds for the base model are

$$\Lambda_{B,\mathrm{core}} = \log(T+1), \qquad \Lambda_{\mathrm{ffn}} = O_{C,R,Z}\left(\sqrt{p} + \frac{2\kappa_C N B_{\mathrm{sel}}}{\delta}\right),$$

with the fixed slots and polynomial FFN allowances of Definition A.4. Closure follows by replacing those slots, as proved in Lemma F.7. No additional score radius is needed: the modification W_a reads the difference directly from the anchor. The arithmetic convention is that of Section 2.2.

Output perturbations under full write-back. To test robustness without changing the training objective, allow an additive perturbation ξ_k to each predicted token before decoding and write-back. It may depend on the instance, model and execution history, but must satisfy $\|\xi_k\|_2 \le \sigma$ on the original predicted TI coordinates. The front end, supplied metadata, stored context and write-back operations remain exact. In the optional counterexample class, the two internal reserved coordinates remain zero at the interface and are not perturbed. No independence or mean-zero assumption is imposed on ξ_k.

Lemma F.1 (Restoration under bounded output errors)**.** *Fix the quantized front end and full write-back, and put $\tau := \min\{\rho, 1/2\} = 1/2$. For an instance X, let $E_k(X)$ and $y_k(X)$ be its stopped teacher prefixes and next-token targets, and write $r_k(F, X) := \|F(E_k(X))_{:,-1} - y_k(X)\|_2$. Set the maximum prediction error to zero when the teacher trace has no anchors. If $\max_{k<L(X)} r_k(F, X) + \sigma < \tau$, then every permitted perturbed rollout of F has exactly the teacher's stored columns and stopping. For any fixed F and $0 \le \sigma < \tau$, its path-failure probability under any such perturbations is at most*

$$1 \wedge \frac{2\mathcal{R}(F)}{(\tau - \sigma)^2}.$$

The same conclusion applies conditional on training when F is learned and the test instance is independent of training.

Proof. Initially, the root and its records equal the teacher's. If the current prefix equals $E_k(X)$, the perturbed output has error at most $r_k(F, X) + \sigma < \tau$. Every discrete block lies within distance $1/2$ of its teacher value. Both nonterminal thoughts lie within ρ of their teacher embeddings and round to them; clipping to the radius-Z ball cannot increase these distances because $Z \ge 1$. Inactive fields and scratch reset, and terminal thoughts become zero. Thus the stored token equals its teacher counterpart. A continuing selection generates the same front-end records from this token, a return uses the same backlink, and the decoded stopping decision is the same. Induction proves the first claim, uniformly over the perturbations. No bound on how F behaves away from teacher prefixes is used.

Consequently, failure can occur only if some prediction error on a teacher prefix is at least $\tau - \sigma$. On that event, $2\ell(F; X) = \sum_{k<L(X)} r_k(F, X)^2 \ge (\tau - \sigma)^2$. Taking expectations of this indicator bound proves the result. The risk is the original sum over a trace, not a step average, so no additional factor B appears. For an empty trace, both failure and the sum are zero. The argument is pointwise in F and therefore also holds conditional on its training data. □

Setting for the sufficient condition. Fix the quantized front end on $\mathcal{E}_H^{(C)}(\bar{\zeta})$ in the stopped problem with budget $B_{\mathrm{sel}} \ge 1$, with $B = 2B_{\mathrm{sel}} + 1$ and $T = (C + 2)(B_{\mathrm{sel}} + 1)$. Use raw feedback on the original causal thought interface with its original targets, $Z \ge 1$ and any fixed attention depth. Let $\mathcal{F}(\mathbf{\Lambda})$ be a pre-sample norm-bounded Transformer class with the input Lipschitz bound L_{in} just derived from Lemma C.3. Let $\mathcal{G}_\alpha \subseteq \mathcal{F}(\mathbf{\Lambda})$ be a minimum-cardinality α-cover under d_∞, with $\mathcal{N}(\alpha) := |\mathcal{G}_\alpha|$ satisfying $\log|\mathcal{G}_\alpha| \le D\log(1 + A_{\mathrm{pred}}/\alpha)$ uniformly over bounded numerical prefixes. Take the quantization radius ρ of Section 5.2 and $N \ge 1$ independent whole instances at confidence $\delta \in (0, 1)$. Set $S_k(L_{\mathrm{in}}) := \sum_{j=0}^{k-1} L_{\mathrm{in}}^j$, with $S_0(L_{\mathrm{in}}) = 0$. Write $L(X) \le B$ for the number of anchors of the stopped teacher trace of X and $y_k(X) := o_{k+1}(X)$ for the target at its k-th anchor prefix $E_k(X)$. For any predeclared $\alpha > 0$, $\bar{\eta} \ge 0$, set

$$r = 3\alpha + \bar{\eta}, \qquad v_N := 1 \wedge \frac{\log\mathcal{N}(\alpha) + \log(1/\delta)}{N} \le 1 \wedge \frac{D\log(1 + A_{\mathrm{pred}}/\alpha) + \log(1/\delta)}{N}.$$

Theorem F.2 (A finite-cover sufficient condition)**.** *With probability at least $1-\delta$ over the N training traces, simultaneously for every $F \in \mathcal{F}(\mathbf{\Lambda})$ satisfying*

$$\widehat{r}_{\max}(F) := \max_{1\le i\le N} \max_{0\le k<L(X_i)} \|F(E_k(X_i))_{:,-1} - y_k(X_i)\|_2 \le \bar{\eta},$$

the fresh-instance probability that any prediction error on a teacher prefix is at least r is at most v_N. If $rS_B(L_{\mathrm{in}}) \le \rho/2$, then outside an event of probability at most v_N its decoded rollout has exactly the teacher's stopped decoded path, the same decoded record data and exact proposal rewards and the same stopping, and every aligned actual thought block is within $\rho/2$ of its teacher value. Exact thought equality is not asserted. At $B_{\mathrm{sel}} = 5H$, its goal probability is at least $1 - H^{-3} - v_N$.

In particular, take

$$\alpha = \bar{\eta} = \eta_* := \frac{\rho}{8S_B(L_{\mathrm{in}})}.$$

Then $r = 4\eta_* = \rho/[2S_B(L_{\mathrm{in}})]$ and

$$v_N = 1 \wedge \frac{\log\mathcal{N}(\rho/[8S_B(L_{\mathrm{in}})]) + \log(1/\delta)}{N} \le 1 \wedge \frac{D\log(1 + 8A_{\mathrm{pred}}S_B(L_{\mathrm{in}})/\rho) + \log(1/\delta)}{N}.$$

For the fixed-depth dense all-trainable class at $B_{\mathrm{sel}} = 5H$ and fixed C under polynomial architecture and norm bounds, this is $1 \wedge O_C((H^3 \log(NH/\delta) + \log(1/\delta))/N)$; on the fixed sparse TI support it is $1 \wedge O_C((H^2 \log(NH/\delta) + \log(1/\delta))/N)$. Every ERM satisfies the dense or sparse guarantee with probability at least $1 - \delta$ over the N training traces when the class contains the exact executor at $t_N = \delta/[2\kappa_C N B_{\mathrm{sel}}]$: use confidence $\delta/2$ for this finite-cover bound, which replaces $\log(1/\delta)$ by $\log(2/\delta)$ in v_N, and the $\delta/2$ sample-margin event proved above. There is no new test-margin term.

Set $\rho_{\mathrm{wb}} = \min\{\rho, 4/5\}$, so $5\rho_{\mathrm{wb}}/8 \le 1/2$ and $5\rho_{\mathrm{wb}}/8 < \rho$. Under the full write-back variant, every discrete block is written back canonically, inactive V, r, h reset, and both thoughts written to nearest node embeddings at nonterminal steps, with zero thoughts at termination. Fix $0 \le \sigma \le \rho_{\mathrm{wb}}/4$ before sampling and allow the output perturbations specified above. Then, with probability at least $1 - \delta$ over the N training traces, every Transformer with $\widehat{r}_{\max}(F) + \sigma \le \rho_{\mathrm{wb}}/4$ has exactly the teacher's stopped tokens and prefixes for every permitted perturbation sequence, except on fresh instances of probability at most

$$1 \wedge \frac{\log \mathcal{N}(\rho_{\mathrm{wb}}/8) + \log(1/\delta)}{N} \le 1 \wedge \frac{D \log(1 + 8A_{\mathrm{pred}}/\rho_{\mathrm{wb}}) + \log(1/\delta)}{N}.$$

The bound has no rollout-amplification factor in B; under the dense polynomial architecture and norm bounds, the failure bound is $\widetilde{O}_C(H^2/N)$. If the class contains the exact executor at $t_N = \delta/[2\kappa_C N B_{\mathrm{sel}}]$, the sample-margin condition above gives zero empirical loss with probability at least $1 - \delta/2$. With probability at least $1 - \delta$ over the N training traces, every ERM therefore satisfies this full-write-back conclusion with $\log(2/\delta)$ in place of $\log(1/\delta)$.

Proof. Choose the finite cover before sampling. For its fixed member G, define

$$A_G(X) = \left\{ \max_{k < L(X)} \|G(E_k(X))_{:,-1} - y_k(X)\|_2 \ge \bar{\eta} + 2\alpha \right\}.$$

If $p_G = \mathbb{P}_X(A_G)$, independence of the whole instances gives exactly $\mathbb{P}\{\mathbb{P}_N A_G = 0\} = (1 - p_G)^N \le e^{-N p_G}$. A union bound shows that, except on an event of probability at most δ over the N training traces, every cover member whose empirical A_G frequency is zero has $p_G \le v_N$ (if the unclipped bound exceeds one the claim is automatic). All these events are measurable for the fixed finite cover and deterministic finite teacher protocol. No independence of prefixes within a trace, ghost sample, or margin assumption is used.

For any F with $\widehat{r}_{\max}(F) \le \bar{\eta}$, choose a cover member G with $d_\infty(F, G) \le \alpha$. On all training prefixes, the prediction error of G is at most $\bar{\eta} + \alpha$, so its empirical A_G frequency is zero. On any fresh prefix, a prediction error of F at least $\bar{\eta} + 3\alpha$ implies a prediction error of G at least $\bar{\eta} + 2\alpha$. This proves the simultaneous prediction-error tail bound. A sample-dependent choice of F and its cover center is permitted because the event is uniform over the fixed finite cover; no measurable selection of the center is needed for the pathwise implication.

Fix a fresh instance on which every prediction error on a teacher prefix is at most r. Compare the raw clipped prefix $\widehat{E}_k$ with the canonical teacher prefix E_k after k Transformer emissions and associated interface writes. Let $M_k = \|\widehat{E}_k - E_k\|_{2,\infty}$ include every predicted coordinate, both thoughts, scratch, records and supplied metadata. Initially, $M_0 = 0$. If the decoded histories agree through step k, their lengths and own positions agree. Direct comparison to the teacher target gives

$$q_{k+1} := \|F(\widehat{E}_k)_{:,-1} - y_k\|_2 \le r + L_{\mathrm{in}} M_k.$$

This uses $F(E_k)$ and does not assume that the fitted F copies any register exactly.

If this bound is at most $\rho/2 < \sqrt{3}/4 < 1/2$, every discrete block decodes to its teacher value. Separate projection of the thoughts onto the radius-Z balls cannot increase their joint error from the teacher thoughts. All remaining coordinates are stored raw, so the full emitted column error is at most q_{k+1}. Inactive V, r, h may be nonzero; their errors remain in this bound.

At a fresh selection before stopping, the clipped thought is within $\rho/2 < \rho$ of its teacher embedding. Quantized lookup therefore gives exactly the same proposal thoughts, embeddings, rewards, validity and valid count. Each candidate record copies a coordinate subvector of the raw emitted column, once per copied field; all other fields and metadata are written exactly by the interface. Its column

error is at most q_{k+1}. Copying the state into C records adds no factor C, since the prefix norm takes a maximum over columns. Thus

$$M_{k+1} \le \max\{M_k, r + L_{\mathrm{in}} M_k\}, \qquad M_0 = 0.$$

Since $S_{k+1}(L_{\mathrm{in}}) = 1 + L_{\mathrm{in}} S_k(L_{\mathrm{in}}) \ge S_k(L_{\mathrm{in}})$ for $L_{\mathrm{in}} \ge 0$, induction gives $M_k \le r S_k(L_{\mathrm{in}})$ and $q_{k+1} \le r S_{k+1}(L_{\mathrm{in}}) \le r S_B(L_{\mathrm{in}})$ for $k < B$. The closure condition supplies the decoding-error bound before each next step. It also keeps every raw column within $\rho/2$ of its aligned teacher column, so every use of the input and cover domain is justified.

At return, the decoded backlink identifies the same earlier column. Its raw source error is already part of M_k. The target remains the teacher's return token, so no learner-source error is duplicated into two reference targets and no $\sqrt{2}$ factor occurs. The same prediction-error bound recovers the decoded return embedding, mask, counters and pointers for arbitrary F. The actual returned thoughts need not coincide.

Acceptance and termination read decoded discrete blocks. The exact external selection counter increments at the same steps. Hence the stopped path and stopping time agree. Terminal thoughts stay within the same error bound of zero but need not be exactly zero. Candidate thoughts are canonical, and all actual source and emitted thoughts obey the full-column bound. Raw nonthought equality is not asserted. The number of forward selections is at most the walk length. Lemma B.8 therefore gives the teacher's goal probability $1 - H^{-3}$ at $B_{\mathrm{sel}} = 5H$; subtracting v_N gives the goal claim. Substitution of $\alpha = \rho/[8S_B(L_{\mathrm{in}})]$ gives the stated logarithm; for the polynomial orders use $\log S_B(L_{\mathrm{in}}) \le \log B + (B-1)\log\max\{1, L_{\mathrm{in}}\}$. The proof uses only the covering-number bound and input Lipschitz bound, so it holds at any fixed attention depth with their corresponding values. The dense and sparse rates use $D = O_C(H^2)$ and $O_C(H)$ respectively, $B = O_C(H)$, constant positive ρ, and the polynomial architecture and norm bounds on A_{pred} and L_{in}.

For ERM, the feasible parameter set is compact and empirical squared loss continuous, so minimizers exist. On the sample-margin event proved above, the exact executor has zero loss. Every minimizer has zero prediction error on every training prefix. Intersect with the finite-cover event at confidence $\delta/2$ and apply the preceding result. The test distribution is unconditioned throughout.

For the full write-back clause, take $\alpha = \rho_{\mathrm{wb}}/8$ and $\bar{\eta} = \rho_{\mathrm{wb}}/4 - \sigma \ge 0$ in the same finite-cover argument. Outside the stated exceptional event, every prediction error on a teacher prefix is strictly less than $r = 3\alpha + \bar{\eta} = 5\rho_{\mathrm{wb}}/8 - \sigma$. Adding any permitted perturbation gives total error strictly below $r + \sigma = 5\rho_{\mathrm{wb}}/8 \le 1/2$ and $r + \sigma < \rho$. Strictness follows from the weak inequality defining A_G; its training prediction error is smaller than that threshold by $\alpha > 0$. Thus even $\rho_{\mathrm{wb}} = 4/5$ stays strictly within every decoding radius. Induct on prediction steps: the root and its records equal the teacher's, and if the prefixes agree, the prediction is evaluated on exactly the teacher prefix, so the input error term $L_{\mathrm{in}} M_k$ is zero. Discrete blocks round to their teacher values and are written back, as are both thought blocks at their teacher embeddings; inactive V, r, h reset, and terminal thoughts are zero. The emitted token therefore equals the teacher's. A fresh selection writes identical candidates and copies the same source register; a return reads the same linked occurrence. Supplied own positions also agree. Thus the next prefixes are exactly the teacher's after full write-back, closing the induction through the common stopping time with the input error term replaced by zero. Substituting $\alpha = \rho_{\mathrm{wb}}/8$ gives the stated bound, with no rollout-amplification factor in B; $D = O_C(H^2)$ gives the dense rate. For the full-write-back ERM claim, on the sample-margin event the exact executor interpolates every training target, so every empirical minimizer has $\widehat{r}_{\max}(F) = 0 \le \rho_{\mathrm{wb}}/4 - \sigma$. Apply this full-write-back cover argument at confidence $\delta/2$ and intersect with the sample-margin event. A union bound gives probability at least $1 - \delta$ over the N training traces and replaces the logarithm by $\log(2/\delta)$, without conditioning the fresh-instance distribution. □

F.2 Full write-back and goal reaching

Definition F.3 (Fixed sparse thought-executor template)**.** Fix the TI layout, one attention stage with $C + 2$ heads, and the separated-stage implementation of F^{C} in Appendix E.1. Fix, before sampling, the union of its weight and bias supports over $0 < t \le 1$; entries outside that support are zero. Every entry on this support is a trainable coefficient. Intersect this linear parameter space with the declared closed norm balls and the linear growth constraints. The positional dictionary, embedding dictionary, and front end are fixed inputs, not trainable tables. The positional reduction inside the initial FFN is included among the trainable coefficients. The dense companion allows every entry in

the same one-stage architecture, with fixed FFN depths $O_C(1)$ and hidden width $O_C(p)$. For Setting 4's counterexample, one may additionally use the fixed FFN slots for W_a of Definition A.4; the cover conclusion below also covers that class. The full write-back result requires only the executor template and no closure condition under W_a.

Corollary F.4 (Covering numbers of the sparse thought-executor class)**.** *The template of Definition F.3 has $m = O_C(p)$ trainable coefficients, and its covering numbers under d_∞ satisfy*

$$\log \mathcal{N}(\alpha) \le D_{\mathrm{sp}} \log(1 + A/\alpha), \qquad D_{\mathrm{sp}} = m = O_C(p), \qquad A = 4\sqrt{m}\,\Lambda_{\max} A_F,$$

where A_F is the parameter Lipschitz constant of Lemma C.3. At $B_{\mathrm{sel}} = 5H$, $p = \Theta_C(H)$, threshold $t_N = \delta/[C(C-1)NB_{\mathrm{sel}}]$, and the norm bounds sufficient to include the exact executor

$$\Lambda_B = \log(T+1), \quad \Lambda_{\mathrm{ffn}} = K_C(\sqrt{p} + t_N^{-1}), \quad \Lambda_V, c_{\mathrm{ffn}} = K_C,$$

for a sufficiently large fixed K_C, the template contains F^{C} at t_N and $D_{\mathrm{sp}} = O_C(H)$, $\log(1 + A) = O_C(\log(NH/\delta))$. The dense companion has $D = O_C(H^2)$ and the same logarithmic order. The optional fixed FFN allowances for W_a retain these dimension orders and, for fixed $a > 0$, polynomial cover constants.

Proof. In the initial FFN, coordinate repairs and signed copies use $O_C(p)$ entries; counter logic has $O_C(1)$ size. Pointer shifts are partial permutations, and their gates and query broadcasts use $O_C(T)$ entries. The positional scalar has the T coefficients $\log(k+1)/\log(T+1)$, not a trainable $T \times T$ table. There are $C+1$ score matrices with the diagonal block $\log(T+1)I_T$ and one zero uniform score, hence $O_C(T)$ score entries. The projection W selects coordinates and each L_j injects them; therefore $V_j = 2L_jW$ has $O_C(p)$ entries, and $V_0 = -\sum_{j=1}^{C+1} L_jW$ also has $O_C(p)$ entries. Output embedding-distance reductions use $O_C(d)$ entries, reward and route comparisons have $O_C(1)$ size, and copies, gates, shifts and constant guards use $O_C(p)$ entries. Reductions and broadcasts remain in separate affine layers, so they create no dense outer product. Signed identity padding adds only $O_C(p)$ entries per layer. There are $O_C(1)$ layers and heads. Thus the total support size is $m = O_C(p)$, including all biases, and is independent of t. Theorem E.8 proves membership of the exact executor at the displayed radii, including linear growth.

For the cover, identify the free entries with $\theta \in [-\Lambda_{\max}, \Lambda_{\max}]^m$. A coordinate grid of mesh $\alpha/(2\sqrt{m}A_F)$ gives parameter distance at most $\alpha/(2A_F)$ in the maximum block operator norm, since each block's operator norm is bounded by the Euclidean norm of all coefficient differences. Replace each occupied grid cell's representative by a feasible point; this at most doubles the distance and preserves the cardinality bound $(1 + 4\sqrt{m}\Lambda_{\max}A_F/\alpha)^m$. Lemma C.3(i),(iii) turns it into a feasible α-cover under d_∞. The parameter set is closed and bounded, hence compact; the linear growth constraint is an intersection of closed inequalities over inputs.

Here, $d = \Theta_C(H)$, $T = (C+2)(5H+1)$ and $B_x = O_C(1)$. All radii are polynomial in NH/δ and depths are fixed, so Lemma C.3's formula makes A_F and A polynomial as well. The dense parameter count is $O_C(p^2)$ and its cover is Lemma C.3(ii). Optional FFN slots for W_a add $O_C(p)$ support entries and one binary architecture choice by Definition A.4 and Lemma F.7; a union of the two covers absorbs $\log 2$ into $D\log(1 + A/\alpha)$ after enlarging D, A by constants. Their polynomial FFN allowances preserve the logarithmic orders. □

Full write-back setting. Use $\mathcal{E}_H^{(C)}(\bar{\zeta})$, $H \ge 2$, fixed $C \ge 2$, $\bar{\zeta} \in (0, 1/2)$, and the quantized front end, which returns only the children and rewards of the node $Q(z)$. Use the stopped TI problem with $B_{\mathrm{sel}} = 5H$, $B = 10H + 1$, $T = (C+2)(5H+1)$, any fixed $Z \ge 1$, and full write-back as in Section 5.2.2; acceptance means the decoded current node is $s^\star$. Assumption 1 holds. Fix $N \ge 1$, $\delta \in (0,1)$, and the sparse or dense class above with radii declared before sampling, containing F^{C} at t_N. Its covering-number bound has constants (D, A) on the bounded-prefix domain of Setting 4. Set $\rho_{\mathrm{wb}} = \min\{\rho, 4/5\}$ and use the maximum training prediction error $\widehat{r}_{\max}$ of Theorem F.2. Fix $0 \le \sigma \le \rho_{\mathrm{wb}}/4$ before sampling for the output perturbations of Lemma F.1. Training minimizes ordinary teacher-forced squared loss on N independent stopped traces. For representation at another $t \in (0, 1]$, use the same executor construction with FFN radius $O_C(\sqrt{p} + t^{-1})$.

Theorem F.5 (Robust full write-back execution, learning, and goal reaching)**.** *In the full write-back setting:*

(a) *F^{C} at threshold t has one attention layer, $C+2$ heads and width $\Theta_C(H)$. On every teacher prefix whose due selection has an eligible reward gap at least t, it emits the teacher's next token exactly before perturbation; returns and termination need no gap. If all selection gaps on an instance's stopped teacher trace are at least t, its full-write-back rollout equals that trace under every permitted output perturbation sequence of norm at most σ. Every stored block is canonical, including interface-written records, and the context has at most $T = O_C(H)$ columns.*

(b) *With probability at least $1-\delta$ over the N training traces, simultaneously every Transformer F in the declared class with $\widehat{r}_{\max}(F)+\sigma \le \rho_{\mathrm{wb}}/4$, and in particular every ERM, has a full-write-back rollout equal to the teacher's stopped trace, with every stored column equal to its teacher counterpart for every permitted perturbation sequence, except on fresh instances of probability at most*

$$v_N := 1 \wedge \frac{\log \mathcal{N}(\rho_{\mathrm{wb}}/8) + \log(2/\delta)}{N} \le 1 \wedge \frac{D\log(1+8A/\rho_{\mathrm{wb}}) + \log(2/\delta)}{N}.$$

Consequently, its decoded path reaches $s^\star$ within $5H$ fresh selections with probability at least $1-H^{-3}-v_N$. Under the displayed polynomial radii, $v_N = \widetilde{O}_C(H/N)$ for the sparse template and $v_N = \widetilde{O}_C(H^2/N)$ for the dense companion.

Proof. For (a), Theorem E.8(a) gives the exact local emission, including inactive zeros. We verify canonical storage through the whole protocol. Initially, both root thoughts are a_{root}, all discrete registers have their prescribed values, scratch is zero, and the supplied position is exact. The root's records contain child embeddings (or zero invalid thoughts and a_{eos}), exact observed rewards, and copies of canonical root fields; their validity, counts, phases and positions are exact. At a selection, the teacher copies a child embedding into both thoughts and writes canonical registers. At a return, both thoughts restore the linked occurrence's embedding, with the correct completed mask, counts and links. Termination writes the fixed token and zero thoughts. Each permitted perturbation has norm at most $\sigma \le \rho_{\mathrm{wb}}/4 < 1/2$ and $\sigma < \rho$, so full write-back restores these exact emissions and resets all inactive fields and scratch, as in Lemma F.1. A continuing fresh state produces the same local records by quantized lookup and copies its now canonical fields; a return or terminal token produces no records. Induction proves equality of every token, record and stop. Here, a canonical reward field in a record is its exact observed reward, not a quantized embedding; no predicted numerical error is retained. The context bound is Definition E.3, and the width follows from $d, T = \Theta_C(H)$.

For (b), the first-entry argument preceding Theorem F.2 gives $\mathbb{P}\{\min_{i\le N} m(X_i) < t_N\} \le N\kappa_C B_{\mathrm{sel}} t_N = \delta/2$. On its complement, (a) makes the exact executor's empirical squared loss zero. Compactness of the feasible parameter set and continuity of empirical loss give existence of ERM; nonnegativity makes every ERM's prediction error zero at every training prefix. Apply the full-write-back clause of Theorem F.2 at confidence $\delta/2$, with $\alpha = \rho_{\mathrm{wb}}/8$ and $\bar{\eta} = \rho_{\mathrm{wb}}/4 - \sigma$. On the intersection of its uniform event and the sample-margin event, which has probability at least $1-\delta$ over the N training traces, simultaneously every Transformer with $\widehat{r}_{\max}(F) + \sigma \le \rho_{\mathrm{wb}}/4$ has all fresh prediction errors on teacher prefixes strictly below $5\rho_{\mathrm{wb}}/8 - \sigma$ except on probability v_N. Every ERM is included because its maximum training error is zero on the sample-margin event. Since $5\rho_{\mathrm{wb}}/8 \le 1/2$ and $5\rho_{\mathrm{wb}}/8 < \rho$, adding any allowed output perturbation still leaves the total error strictly below every discrete decoding radius. Lemma F.1 therefore gives equality of every stored column, including records, and the entire stopped trace for every such Transformer. This uses no fresh margin restriction and no additional test-margin probability.

Under this front end, the canonical teacher is reward-ordered DFS on the original tree: selections visit uncompleted children in decreasing reward order and returns mark exactly the completed child. Couple it to the walk in Definition B.4. Each downward walk move is one fresh selection; each upward move is one return; candidate record writes are neither. Therefore the number of fresh selections through goal time τ_G is at most τ_G. On $\{\tau_G \le 5H\}$ the budget $B_{\mathrm{sel}} = 5H$ cannot stop the teacher before its goal: strictly before that goal it has used fewer than $5H$ selections. If the goal is the last allowed selection, its decoded acceptance is still tested on that emission before further records are suppressed. Termination cannot precede the goal in this finite DFS traversal. The matched rollout has the same acceptance and selection count. Lemma B.8 and a union bound, with

no independence requirement, now give

$$\mathbb{P}_X(\text{rollout reaches } s^\star \text{ within } 5H \text{ fresh selections}) \geq \mathbb{P}_X(\tau_G \leq 5H) - \mathbb{P}_X(\text{trace mismatch}) \geq 1 - H^{-3} - v_N.$$

Finally, $\rho_{\mathrm{wb}} > 1/2$ is bounded away from zero. Corollary F.4 gives, for the sparse and dense classes respectively,

$$v_N \leq 1 \wedge O_C\left(\frac{H\log(NH/\delta) + \log(2/\delta)}{N}\right), \qquad v_N \leq 1 \wedge O_C\left(\frac{H^2\log(NH/\delta) + \log(2/\delta)}{N}\right).$$

These are the stated rates. The cover bounds an instance-level maximum prediction error event, so the number of anchors contributes no factor outside the logarithm. □

F.3 A COUNTEREXAMPLE IN THE ORIGINAL THOUGHT INTERFACE

Use the quantized front end on $\mathcal{E}_H^{(C)}(\bar{\zeta})$, Definition E.1's original layout, raw feedback, $Z \geq 1 + a$ and $B_{\mathrm{sel}} \geq H$, with a fixed below. Put $B = 2B_{\mathrm{sel}} + 1$ and $T = (C+2)(B_{\mathrm{sel}} + 1)$. The root thought is its node embedding. Every column of every queried teacher prefix satisfies $z_{\mathrm{src}} = U$. At a fresh state, this follows from the embedding-valued proposal; candidate records copy both fields; a teacher return restores the linked parent's equal fields. Invalid leaf records also copy these fields, although their proposal thought is zero. The stopped teacher never emits a terminal token: the goal is a leaf and DFS finds it before exhausting the root, unless the selection budget stops the trace first. Thus at every queried anchor included in the loss the two output targets are the same node embedding. Terminal tokens, whose source is zero and whose embedding is a_{eos}, are neither queried nor targets in these traces.

Definition F.6 (The modification W_a)**.** Fix $a \in (0, \min\{1, \rho/2\}]$, gain 2, and

$$S_k(2) = 2^k - 1, \qquad u = \frac{a}{S_{H-1}(2)}, \qquad K = [-d^{-1/2}, d^{-1/2}]^d.$$

For $G = F_{G_0,b}$ in Definition A.4, the modification W_a sets its fixed slots to $b = 1$. At the last input column x, preserve the source–embedding difference $q = (z_{\mathrm{src}}(x) - U(x))_1$ before any base operation. Both fields are already in every anchor: the last candidate record copies them from its fresh occurrence, and a return anchor carries its own. Let $y = G_0(E)_{:,-1}$ be the base Transformer's raw output and set

$$m_0 = (y_{z_{\mathrm{src}}} + y_U)/2, \qquad m = P_K(m_0), \qquad \bar{q} = \mathrm{clip}_{[-a,a]}(q), \qquad d^+ = 2\bar{q} + u\kappa,$$
$$z_{\mathrm{src}}^+ = m + \tfrac{1}{2}d^+ e_1, \qquad U^+ = m - \tfrac{1}{2}d^+ e_1, \qquad g(q) = f_{a/2,a}(|q|).$$

Here, $\kappa = 1$ on every actual input column, since $\pi = e_i$; P_K is coordinatewise clipping and f is the ramp of Lemma A.6(ii). Thus $g = 0$ on $|q| \leq a/2$ and $g = 1$ on $|q| \geq a$; the band is immaterial on the analyzed descent. Use $M_{\mathrm{core}} = \max\{2, B_0\}$ declared in Definition A.4. Feed $\widetilde{U} = \mathrm{clip}_{[-M_{\mathrm{core}}, M_{\mathrm{core}}]}(U^+)$ coordinatewise to the bounded overwrite gadget:

$$U_{W_a(G)} = \Gamma_{M_{\mathrm{core}}}(1 - g(q), \widetilde{U}) + \Gamma_{M_{\mathrm{core}}}(g(q), a_{\mathrm{root}}). \tag{F.1}$$

The output source is z_{src}^+; z and every other original output block, including the base Transformer's scratch, remain as G_0 (and G) emits them. Both reserved coordinates are cleared after use. These formulas are used at every queried anchor; outputs at unqueried columns never enter the loss or the interface. For $|q| \leq a$, the difference update is $d^+ = 2q + u$; when $|q| \leq a/2$, the gate is zero and the emitted difference is exactly d^+. The analysed descent moves from this region directly to the trigger $q = a$. The clip in equation F.1 bounds the gadget input on all numerical prefixes; at gate one, the overwrite is exact regardless of the base Transformer. No route repair is required for departure: an incorrect route already differs from the teacher's, and a correct selection route with $U = a_{\mathrm{root}}$ also differs.

Write $\mathsf{S} \circ G$ for the map replacing G's actual (z_{src}, U) outputs by their projected mean, with every other output unchanged. At teacher anchors, this mean equals m for both choices of b; this fact, proved below, is what the loss identity uses. Away from teacher prefixes, $W_a(F_{G_0,1}) = F_{G_0,1}$ by slot replacement, not by applying the displayed pair operation twice. The infimum below is over this single declared class. The modification W_a uses the existing anchor fields and two reserved internal coordinates that remain zero at the interface. No class membership of $\mathsf{S} \circ G$ is needed.

Lemma F.7 (Pair symmetrization, loss identity, and resources)**.** *The class of Definition A.4 contains the exact executor and is closed under W_a with the same declared architecture constants. Its covering numbers, input Lipschitz bound and output envelope are bounded by those of the ambient architecture with token width $p' = p + 2$ and the declared architecture and norm bounds. For every member G and instance X, with $L(X) \le B$ queried anchors,*

$$\ell(W_a(G); X) = \ell(\mathsf{S} \circ G; X) + \frac{L(X)u^2}{4} \le \ell(G; X) + \frac{L(X)u^2}{4}. \tag{F.2}$$

The equality holds anchor by anchor with penalty $u^2/4$. It holds for population and empirical averages. The coefficient of the incoming q contributes zero on every teacher prefix; the positive initial offset is visible in the loss.

At fixed base depth, the modification W_a uses $O_C(p')$ additional FFN nonzeros, $O_C(p')$ hidden width and $O_C(1)$ additional affine stages, within the explicit allowances of Definition A.4. A sufficient new-layer weight and bias radius is

$$K_C\big((1 + M_{\mathrm{core}})\sqrt{p'} + a^{-1}\big),$$

for a constant K_C independent of H and u^{-1}. The containing FFN radius is the maximum of this and the base model's norm bound (enlarged by a constant for identity channels). A sufficient per-FFN linear growth allowance is $K_C(1+c_0)(1+a^{-1})\sqrt{p'}$. Attention depth, head count, score radius and value radius need not increase. These are polynomial costs for fixed $a > 0$ and polynomial architecture and norm bounds for the base model; the exponentially small initial offset requires no large weight.

Proof. For any pair (y_{src}, y_U), set $m_0 = (y_{\mathrm{src}} + y_U)/2$ and $d_0 = y_{\mathrm{src}} - y_U$. For a common target $v \in K$, expansion cancels the cross terms:

$$\tfrac{1}{2}\big(\|y_{\mathrm{src}} - v\|^2 + \|y_U - v\|^2\big) = \|m_0 - v\|^2 + \tfrac{1}{4}\|d_0\|^2.$$

Coordinate projection gives $\|P_K(m_0) - v\| \le \|m_0 - v\|$: in each coordinate clipping an exterior value toward the interval cannot increase its distance from a point in that interval. Hence pair symmetrization never increases loss, for either $b = 0$ or $b = 1$. On every teacher anchor, $q = 0$, $g(q) = 0$ and $d^+ = u$, so the new pair's contribution is exactly $\|m - v\|^2 + u^2/4$. The preliminary clip of U^+ is inactive there. Indeed $u \le a \le \rho/2 < \sqrt{3}/4$, so $\|U^+\|_\infty \le 1 + a/2 < 2 \le M_{\mathrm{core}}$. For $G = F_{G_0,0}$, this is the claimed identity directly. For $G = F_{G_0,1}$, its existing pair is already $(m + ue_1/2, m - ue_1/2)$. Its projected mean is $P_K(m) = m$, so $\mathsf{S} \circ G$ has the same pair (m, m) as $\mathsf{S} \circ F_{G_0,0}$, while slot replacement leaves $W_a(G) = G$. This proves the identity for every member, not just for a base Transformer. All other outputs are unchanged. The two reserved targets and emitted coordinates are zero, so their squared loss is zero, anchor by anchor. No coordinate is silently omitted from the original risk accounting.

For implementation, the first fixed slot writes q and κ into the two reserved token coordinates before the base Transformer's initial FFN. Every base FFN passes them unchanged, using their positive and negative parts in hidden layers; its original outputs ignore them. The zero rows and columns of the attention maps prevent any read or write there, and attention skip connections carry them to the output slot. This applies to every column, in particular the last anchor column. Means, scalar and coordinate clips, absolute value, the ramp, and the bounded overwrite are exact ReLU formulas from Lemma A.6; for example $\mathrm{clip}_{[-b,b]}(v) = \sigma(v+b) - \sigma(v-b) - b$. They have a fixed number of successive stages. Carry every original output coordinate through these stages, so even the scratch outputs of an arbitrary base Transformer are preserved. Each coordinate uses $O(1)$ nonzeros; only $O_C(1)$ scalars are broadcast. With reductions and broadcasts in separate layers, there are $O_C(p')$ added nonzeros in total at fixed depth. The capture and guard need at most four affine stages; the mean and clips, difference update and ramp, and bounded overwrite and clearing need at most twelve further stages, including signed identity transport. The allowances of 32 extra stages and $64p'$ hidden units therefore suffice, also for identity padding. Coefficients have magnitude $O(1 + M_{\mathrm{core}})$ apart from the scalar ramp slope $2/a$. Summing squared entries gives the stated weight and bias bound. In particular, there is no coefficient $1/u$.

The maps must also vanish at zero in the class convention. For this purpose alone carry the constant guard $\kappa = s_{\mathrm{bit}}(\sum_i \pi_i)$ from the original supplied own-position block, and implement the initial

offset as $u\kappa$. Every actual input column already has $\pi = e_i$, so $\kappa = 1$ even on raw return anchors; no position index or new metadata is used. At zero, $\kappa = 0$, and $|\kappa| \le 2\sqrt{T}\|\pi\|_2$. All the displayed anchor formulas therefore hold exactly. The projected center and clips satisfy linear growth bounds, $|g(q)| \le 2|q|/a$, and $\|\Gamma_M(g, v)\|_2 \le 2\|v\|_2$. For $g \in [0, 1]$, $M \ge |v|$ and a scalar constant v, $|\Gamma_M(g, v)| = \max\{|v| - M(1 - g), 0\} \le g|v|$. Hence the fixed root branch has norm at most $g(q)$, since its embedding has norm one and $M_{\text{core}} \ge 1$. Thus the whole output map has linear growth coefficient $O_C(1 + a^{-1})$ in its base Transformer's outputs and reserved coordinates, and the input guard costs at most $O(\sqrt{p'})$. Together with identity extensions of the base Transformer, this proves the stated allowance. All guard and difference update wires are internal and are cleared in the emitted token.

Embed F^{C} as $F_{F^{\mathrm{C}},0}$. Its original calculation is unchanged and its reserved coordinates stay zero, so it belongs to the declared class. For any $G = F_{G_0,b}$, replacement gives $W_a(G) = F_{G_0,1}$ with exactly the same base parameters and fixed slots. The bounds just proved apply uniformly over the base class, and $W_a(W_a(G)) = W_a(G)$. No layer, parameter radius or support is consumed on reapplication. This proves closure of the declared class; the slack verifies its realization in the declared architecture, not closure of an unrestricted norm ball.

Finally, apply Lemmas C.1 and C.3 to that architecture with $p' = p + 2$. Zero padding preserves input and target norms. In particular, its envelope is $B_{\text{out}} = c_{\text{ffn}}^{L_{\text{att}}+1}(1 + J\Lambda_V)^{L_{\text{att}}}B_x$, and its covering numbers satisfy Lemma C.3 with

$$D = O\big(L_{\text{att}}J(p')^2 + (L_{\text{att}} + 1)L_{\text{ffn}}(W_{\text{ffn}}^2 + p'W_{\text{ffn}} + p')\big).$$

The same parameter partition proof, choosing a representative only from cells meeting the declared subclass, gives a feasible cover inside it, with the same D and covering-number constant up to a fixed factor. Fixed sparse support uses its declared nonzero count. The input Lipschitz bound is precisely the ambient bound in the input-stability paragraph, at these same widths, depths and radii. At $B_{\text{sel}} = 5H$, base width $O_C(H)$, fixed depths and fixed C, a, polynomial base radii still give polynomial A, L_{in}, dense $D = O_C(H^2)$ and sparse $D = O_C(H)$. The sufficient-condition and write-back arguments use only these bounds and zero reserved coordinates at the interface; Theorem 4(a), including its ERM clause, and the full-write-back guarantee of Theorem 5 therefore apply unchanged. □

Theorem F.8 (A counterexample with exponentially small excess risk)**.** *In the setting of Definition F.6, use the pre-sample class $\mathcal{F}(\mathbf{\Lambda})$ of Definition A.4. For every $\epsilon > 0$, it contains F^{bad} with*

$$\mathcal{R}(F^{\text{bad}}) \le \inf_{F \in \mathcal{F}(\mathbf{\Lambda})} \mathcal{R}(F) + \epsilon + \frac{(B + 1)a^2}{4(2^{H-1} - 1)^2}, \tag{F.3}$$

whose rollout under raw feedback leaves the teacher's decoded trace by the first selection of a depth-H node, on every instance whose stopped teacher trace reaches depth H. In particular, this applies to every instance on which the teacher reaches the goal. Earlier incorrect decoded actions, malformed continuations and premature stops are already departures; a difference in numerical thoughts or proposals alone is not counted as decoded departure. For a trace stopped by budget before depth H, only the alternative of earlier departure or agreement until that stop is asserted. In the complete-tree setting with $B_{\text{sel}} \ge H$, every teacher trace in fact reaches depth H at its Hth fresh selection.

The modification W_a satisfies the exact loss identity equation F.2. For $B = \Theta_C(H)$ and fixed $a > 0$, its additive allowance in equation F.3 is $\Theta_C(H4^{-H}) = e^{-\Theta(H)}$, with polynomial radii whenever the base model's norm bounds are polynomial. Taking ϵ at this order gives exponential excess over the same class's infimum. This does not assert exponentially small absolute risk or failure of an interpolating ERM.

Proof. Choose an ϵ-near-minimizer $G \in \mathcal{F}(\mathbf{\Lambda})$ for population risk and put $F^{\text{bad}} = W_a(G)$. Closure gives membership in the same class. Lemma F.7 gives the sharper bound with $\mathbb{E}L(X)u^2/4 \le Bu^2/4$; using the conservative $(B + 1)u^2/4$ proves equation F.3. No comparison of infima of different classes, or membership of the symmetrized map, enters this argument.

Fix an instance and suppose there has been no decoded departure before the teacher's first depth-H selection. The complete tree has no leaf above depth H, so its first H selections are an uninterrupted

descent. An agreeing rollout must emit the same fresh-selection actions and discrete registers; hence it too has no intervening return or stop. Its front-end proposals may differ, because the arbitrary base Transformer's z output need not be canonical. This does not affect the argument: every last candidate record copies the occurrence's raw source and current-embedding fields, irrespective of its proposals. These copied fields, not a lookup result or the base Transformer's next prediction, supply q.

Let q_k be this source–embedding difference in the state after the kth selection, with $q_0 = 0$ at the root. Inductively, before the override,

$$q_{k+1} = 2q_k + u \quad (0 \le k < H-1), \qquad q_k = uS_k(2) \quad (0 \le k \le H-1).$$

To justify all clips and gates in this induction, for $k \le H-2$,

$$0 \le q_k \le a\frac{2^{H-2}-1}{2^{H-1}-1} < a/2, \qquad 0 \le q_{k+1} \le a.$$

Thus $\bar{q} = q_k$ and $g(q_k) = 0$. Since $\|m\|_2 \le 1$, the emitted source has norm at most $1 + a/2 \le Z$, so the interface thought clip is inactive on it. Also $\|U^+\|_\infty \le 1 + a/2 < 2 \le M_{\text{core}}$, so its preliminary clip is inactive. The emitted difference is exactly $(2q_k + u)e_1$, independent of the center, the base Transformer's other outputs, and any clipping of z. Candidate records preserve this difference exactly, closing the induction. There are no returns in the interval being proved; no exact-copy property of an arbitrary base Transformer on returns is assumed.

The input at selection H is therefore $q_{H-1} = a$. At this trigger,

$$\|z^+_{\text{src}}\|_2 \le 1 + a + u/2 \le 1 + 3a/2 < 1 + 3\sqrt{3}/8 < 2.$$

Thus the trigger source is unclipped when $Z = 2$. For smaller allowed $Z \ge 1 + a$, it may be clipped at this step; no further recurrence is needed, and thought clipping cannot change the node-embedding override. The ramp is exactly one. The clipped overwrite input obeys $\|\widetilde{U}\|_\infty \le M_{\text{core}}$, so equation F.1 suppresses it and writes a_{root} exactly, for every base Transformer. Raw feedback fixes this embedding and enforces no child-legality correction. If other emitted discrete blocks still describe the teacher's selection, its node embedding now differs, because the root is not an eligible child of the teacher's depth-$(H-1)$ node. If those blocks differ, departure is already established. This comparison precedes any acceptance or budget test caused by the output, including when the teacher's leaf is the goal or its Hth selection exhausts the budget.

A teacher trace stopped before the triggering selection has no later emission to compare: it may agree up to its stop, and the conditional claim imposes no failure after that stop. Under the present $B_{\text{sel}} \ge H$ hypothesis, this case is empty, since a goal cannot occur above depth H and the initial descent takes exactly H selections. The conditional formulation states the stopping convention explicitly without treating numerical proposal drift as path departure.

Finally, for $H \ge 2$,

$$2^{H-2} \le 2^{H-1} - 1 \le 2^{H-1}, \qquad (B+1)a^2 4^{-H} \le \frac{(B+1)a^2}{4S_{H-1}(2)^2} \le 4(B+1)a^2 4^{-H}.$$

This proves the exponential order; the resource lemma proves the polynomial radius assertion. Every teacher trace has at least one anchor, so equation F.2 gives positive empirical loss for this modified model when $u > 0$. It cannot be an ERM on an interpolable sample. Near-optimal teacher-forced risk and the sufficient condition on the maximum prediction error are different conditions; the proof asserts no matching lower bound for the sufficient tolerance $\rho/[8S_B(L_{\text{in}})]$. □

Corollary F.9 (Counterexample based on the exact executor)**.** *Assume $F^{\text{C}}_t \in \mathcal{F}(\mathbf{\Lambda})$. Take $G = F^{\text{C}}_t$ at threshold $t \in (0, 1]$ and $F^{\text{clean}} = W_a(F^{\text{C}}_t)$ in the declared class. Let G_t be the event that all first H selection gaps of the teacher's initial descent are at least t. Then $\mathbb{P}(G^c_t) \le CHt/2$. On G_t, the decoded path agrees with the teacher through selection $H-1$, the emitted z and all front-end proposal thoughts, embeddings and rewards are canonical, and at selection H, the node embedding is exactly a_{root}. The source and copied raw embeddings need not equal the teacher's. The offending emission violates closed-loop correctness and its on-policy copy loss is at least $1 - \varepsilon > 1/2$. No assertion about eventual goal failure after departure is made. Let $B_{\text{out},C}$ and $B_{y,C}$ be the output*

and target bounds of Lemma C.1 specialized to the exact executor and its teacher prefixes. Its fixed C, R, Z and constant linear growth bounds give $\bar{b}_{\mathrm{TI},C} := (B_{\mathrm{out},C} + B_{y,C})^2/2 = O_C(1)$. Then

$$\mathcal{R}(F^{\mathrm{clean}}) \le \bar{b}_{\mathrm{TI},C}\kappa_C B_{\mathrm{sel}} t + \frac{Ba^2}{4S_{H-1}(2)^2}.$$

At $B_{\mathrm{sel}} = \Theta_C(H)$ and $t = H^{-r}$, $r > 1$, this is $O_C(H^{1-r}) + O_C(H4^{-H})$ with polynomial architecture and norm bounds.

Proof. Condition on the visited-node history of the initial descent. Each new outgoing block is fresh, so the top-gap bound of Lemma B.3 and a union bound give $CHt/2$. Induct on $k < H$. At each selection, the exact executor's repaired registers and addressed reads give the same winning canonical proposal v. It emits $y_z = y_{z_{\mathrm{src}}} = y_U = v$, hence the modification's center is v, its z is v, and its source and embedding are $v \pm q_k e_1/2$. Their deviations have norm at most $a/2 \le \rho/4 < \eta$. To verify the last strict inequality uniformly, choose four distinct node embeddings, which exist for $H \ge 2$. Expanding the squared norm of their sum shows that some pair has inner product at least $-1/3$, so $\rho \le \sqrt{2/3}$. Hence $\rho/4 \le 1/\sqrt{24} < 3/14 < \eta$. The parent-embedding copy has the same bound; all other discrete registers are canonical. Thus the next prefix is in the scope of Theorem E.8(c), and decoding and quantized lookup select precisely the teacher's node and proposals. Candidate copies of the source and embeddings retain their small raw deviations; these are not claimed to equal the teacher's raw records. This closes the induction and proves closed-loop correctness through $H - 1$. At the next query, the input remains admitted, but the modification W_a writes the root instead of the selected proposal embedding v. Their half squared distance is $\frac{1}{2}\|a_{\mathrm{root}} - v\|^2 = 1 - \langle a_{\mathrm{root}}, v\rangle \ge 1 - \varepsilon$. This is one block's contribution to the on-policy copy loss, evaluated before the resulting scope exit or stop.

On teacher prefixes, the exact executor loses only at sub-threshold selections, with per-step loss at most $\bar{b}_{\mathrm{TI},C}$. The first-entry adjacent-gap argument in the trace-margin paragraph bounds their expected number by $\kappa_C B_{\mathrm{sel}} t$. Pair symmetrization can only reduce this loss; equation F.2 adds at most $Bu^2/4$. Substituting $t = H^{-r}$ gives the final order. The base Transformer's score radius is still $\log(T+1)$ and its FFN radius is $O_C(\sqrt{p} + t^{-1})$; the modification W_a adds only the polynomial allowance already proved. □

Corollary F.10 (The same counterexample under full write-back)**.** *Use Setting 4's counterexample class on the complete tree with $H \ge 2$ and $B_{\mathrm{sel}} \ge H$, including the exact executor at $t_N = \delta/(2\kappa_C N B_{\mathrm{sel}})$. Let F^{bad} be the model of Theorem F.8 for $\epsilon > 0$. Keep its parameters, front end and environment fixed, and change only the storage protocol to full write-back. For every output perturbation strategy of Lemma F.1 with $0 \le \sigma < 1/2$, the probability of departure from the teacher's stopped trace is at most*

$$1 \wedge \frac{\bar{b}_{\mathrm{TI},C}\delta/N + 2\epsilon + (B+1)a^2/[2S_{H-1}(2)^2]}{(1/2-\sigma)^2}.$$

For fixed a and σ bounded away from $1/2$, and $B = \Theta_C(H)$, this is $O_C(\delta/N + \epsilon + H4^{-H})$. In contrast, the same model with unperturbed raw feedback departs by the first depth-H selection on every instance. At $B_{\mathrm{sel}} = 5H$, full-write-back goal success is at least $1 - H^{-3}$ minus the displayed failure bound.

Proof. Write $R_* := \inf_{F\in\mathcal{F}} \mathcal{R}(F)$ for the infimum over this same declared class and $u = a/S_{H-1}(2)$. Lemma F.1 and Theorem F.8 give, for the same F^{bad}, failure probability at most

$$1 \wedge \frac{2R_* + 2\epsilon + (B+1)u^2/2}{(1/2-\sigma)^2}.$$

It remains to bound R_* without assuming it is zero. On teacher prefixes, $F^{\mathrm{C}}_{t_N}$ is exact except at selections whose eligible reward gap is below t_N. Each such selection contributes at most $\bar{b}_{\mathrm{TI},C}$ to the half squared loss, and their expected number is at most $\kappa_C B_{\mathrm{sel}} t_N$, by the trace-margin argument and the output bound used in Corollary F.9. Since the exact executor belongs to this class,

$$R_* \le \mathcal{R}(F^{\mathrm{C}}_{t_N}) \le \bar{b}_{\mathrm{TI},C}\kappa_C B_{\mathrm{sel}} t_N = \frac{\bar{b}_{\mathrm{TI},C}\delta}{2N}.$$

Substitution proves the bound. The exponential term follows from $S_{H-1}(2) = 2^{H-1} - 1$ and $B = \Theta_C(H)$. The raw-feedback assertion is exactly the existing counterexample: the complete-tree teacher reaches depth H in its first H selections, before any possible earlier budget stop. For the goal conclusion, combine trace agreement with Lemma B.8 by a union bound.

This argument applies to the near-infimum counterexample itself, not only to the exact-executor witness. It does not require that F^{bad} satisfy the empirical prediction-error condition of Theorem 5, and does not identify it as an ERM. The parameters N, δ here specify the predeclared class and exact-executor threshold; no additional event involving the training sample is used in this population-risk comparison. Perturbations leave the two reserved internal coordinates of W_a zero, as stipulated above. □

G Experimental details

We report one teacher-forcing experiment with two feedback protocols. At each depth $H \in \{4, 6, 8, 10\}$, we train ten models with seeds 0–9 and evaluate each on 2048 shared test environments.

G.1 Setup

Environment and embeddings. Training, validation and testing use the binary environment of Setting 1 at $\bar{\zeta} = 0.2$. Writing $L_2 = \lfloor \log_2 H \rfloor$ and $r_i = H - i + 1$, the misranking probabilities are $\zeta_i = 0.2$ for $r_i \le L_2$ and $\zeta_i = H^{-3} 2^{L_2 - r_i}$ otherwise. We set $\sigma_i = [\sqrt{2}\,\bar{\Phi}^{-1}(\zeta_i)]^{-1}$ to obtain misranking probability ζ_i, with reward transform $\psi_i(y) = \Phi(y/\sigma_i) + \Phi((y-1)/\sigma_i) - 1$. Slots follow child order, not reward order; latent preferences and the goal never enter model inputs. These calculations use float64 temporaries, with rewards and tokens stored in float32. Endpoint rewards are moved inside $(-1, 1)$; exact sibling ties raise an assertion.

For each (H, seed), sequential rejection constructs a fixed dictionary in $\{\pm 1/\sqrt{128}\}^{128}$ with absolute pairwise inner products at most 0.35, including separate null-parent and EOS embeddings. The front end reads the nearest node embedding $Q(z)$ and returns that node's child embeddings as candidate thoughts, with rewards and validity bits. Leaves have two invalid slots with zero thoughts and rewards and EOS embeddings. Nearest-embedding ties select the lowest index. The node-only radius ρ is reported in Table 3.

Interface and feedback. The token sequence implements Definitions E.1, E.2 and E.3. The implementation stores the two thoughts, node and parent embeddings, candidate fields, completed-slot mask, counts and backlinks. A separate candidate block duplicates the candidate record's thought; constant and zero scratch fields are omitted. Own position is supplied and excluded from prediction and loss. Fresh occurrences append two candidate records; the second is an anchor. A return is itself an anchor, restores the linked parent visit and updates its completed-child information without generating new records.

A decoded view dispatches actions, resolves addresses and tests stopping; it does not repair illegal parent–child choices. Raw feedback clips only the two thoughts to radius $Z = 2$ and stores other predictions raw. Full write-back decodes the discrete fields, clears inactive fields, rounds both nonterminal thoughts to node embeddings and zeros terminal thoughts. The fixed front end uses $Q(z)$ only, not the decoded address to correct a thought. The teacher and learned rollouts share the interface and stopping rules: stop at the goal, termination or the $5H$-th fresh selection, before adding further records. The root and returns do not count as fresh selections; $B = 10H + 1$ and $T = 4(5H + 1)$. Malformed predictions are recorded as failures. Reaching the B-prediction limit before another stopping condition is also recorded as failure.

Model and training. An initial MLP maps input $\to 512 \to 512 \to 384$ with two ReLUs. One causal softmax attention layer has four learned Q/K/V heads of dimension 64, a learned 256-to-384 output projection, and a skip connection. The output MLP maps $384 \to 512 \to 512 \to$ predicted blocks with two ReLUs and a linear readout. Every linear map has a bias. There is no layer normalization or dropout. Input and predicted widths are $785 + 4T$ and $785 + 3T$; parameter count is $2{,}118{,}929 + 3587T$ (Table 1). Each run uses $N = 4096$ fresh training instances and 128 held-out validation instances. A batch samples 16 distinct training instances and includes every

anchor. The loss sums squared error over all predicted coordinates and averages over anchors. Targets use the teacher's canonical thoughts. AdamW uses learning rate 10^{-3}, weight decay 0.01, 500 warmup steps, cosine decay to zero over $20{,}000$ steps, and gradient clipping at norm 1. All model arithmetic is float32. Random streams for embeddings, training, validation, batches, and evaluation are separate; deterministic algorithms are enabled. Every 500 steps, validation loss and the maximum raw Euclidean token prediction error over all training anchors are measured. The selected checkpoint is the latest evaluated iterate within 1.10 times the running minimum validation loss. No rollout metric enters selection. The reported $\widehat{\eta}$ belongs to this checkpoint and includes every predicted block before projection.

Table 1: Model size by depth; T is the context bound $4(5H + 1)$.

H	T	Input / predicted width	Parameters
4	84	1121 / 1037	2,420,237
6	124	1281 / 1157	2,563,717
8	164	1441 / 1277	2,707,197
10	204	1601 / 1397	2,850,677

The recorded training and validation losses are mean squared errors per anchor, not the trace-summed risk with factor $1/2$ in Section 2.2. This optimizer and parameterization do not implement ERM in the theorem's prescribed norm-bounded class.

Evaluation and diagnostics. Each selected model is evaluated on 2048 environments at evaluation seed 20260923. For a fixed depth, these environments are shared across training seeds and feedback protocols. Path success compares all decoded discrete emissions, lengths and stopping reasons with the teacher; closed-loop correctness compares each decision with the rule on the rollout's own prefix. Goal success only requires reaching the decoded goal before stopping. Path success and closed-loop correctness coincide on all reported examples. The teacher reaches the goal on $2045/2048$ environments at $H = 4$, $2047/2048$ at $H = 6$, and all 2048 at $H = 8, 10$.

The embedding-error monitor uses $3(1 - 0.35)/7$, as in Setting 3, and the gap monitor uses $\delta/(2NB_{\mathrm{sel}})$ with $\delta = 0.05$. A monitor violation is recorded but does not stop or correct the rollout. The logged field `first_cell_exit` tests distance from the teacher embedding against ρ; it records leaving a sufficient safe radius, not necessarily changing the nearest embedding. Drift is measured only while the compared path remains aligned; rule diagnostics stop at the first incorrect decision. Missing later diagnostics are not zero errors.

G.2 Results and uncertainty

Table 2 reports every trained seed. The full-write-back advantage is small at $H = 4, 6, 8$ and larger at $H = 10$. At $H = 10$, raw path success has seed standard deviation 7.91 percentage points, compared with 2.13 for full write-back.

Table 4 gives paired differences. We average each environment's difference across the ten models, then bootstrap the 2048 environment identities 5000 times. The intervals condition on the fitted models; they do not quantify uncertainty over training seeds. Rescued counts failures under raw feedback that become successes under full write-back; lost counts the reverse. Both refer to $20{,}480$ model–environment pairs per depth.

At $H = 10$, raw feedback ends with $18{,}336$ goal stops, 2068 selection-budget stops and 76 prediction-limit stops over $20{,}480$ model–environment evaluations. Full write-back gives $20{,}102$, 360 and 18, respectively. No non-finite or malformed-prediction stop occurs in these forty runs. Under full write-back, 3315 evaluations still reach the goal after departing from the teacher's path, underscoring the distinction between the two success notions.

The maximum training errors exceed the full-write-back tolerance by factors of 5.43 to 17.07. The sufficient raw-feedback tolerance is no larger, since $S_B(L_{\mathrm{in}}) \geq 1$. Neither sufficient condition therefore holds for these checkpoints. The experiments show a storage-dependent execution difference, not necessity of quantization, ERM learning, or an empirical verification of exponential error growth.

Table 2: All forty runs: success rates (%). Raw is raw feedback; WB is full write-back. Each row contains the same 2048 paired test environments for its depth.

H	Seed	Raw path	WB path	Raw goal	WB goal
4	0	97.71	97.71	99.56	99.56
4	1	97.31	97.41	99.61	99.66
4	2	96.97	97.12	99.46	99.61
4	3	97.31	97.51	99.41	99.66
4	4	98.00	98.05	99.66	99.71
4	5	97.51	97.56	99.46	99.61
4	6	97.66	97.71	99.61	99.61
4	7	97.80	98.00	99.61	99.85
4	8	97.41	97.46	99.71	99.76
4	9	97.71	97.80	99.66	99.71
6	0	97.17	97.17	99.85	99.85
6	1	96.78	96.78	99.85	99.85
6	2	96.78	96.83	99.76	99.85
6	3	96.44	96.44	99.90	99.85
6	4	96.88	97.02	99.46	99.71
6	5	96.53	96.63	99.71	99.80
6	6	97.07	97.31	99.71	99.80
6	7	97.22	97.31	99.61	99.71
6	8	96.34	96.39	99.85	99.90
6	9	96.53	96.63	99.85	99.90
8	0	87.99	88.48	99.41	99.76
8	1	90.67	91.06	99.76	99.95
8	2	88.67	88.92	99.71	99.85
8	3	89.06	90.09	99.76	99.90
8	4	87.89	88.28	99.66	99.90
8	5	89.26	89.70	99.41	99.80
8	6	88.96	89.55	99.61	99.85
8	7	89.89	90.82	99.07	99.85
8	8	88.62	89.16	99.56	99.90
8	9	89.50	90.04	99.51	99.85
10	0	76.76	83.20	94.53	98.88
10	1	68.75	80.76	88.72	97.71
10	2	83.01	84.81	97.17	97.95
10	3	64.60	81.15	85.74	97.90
10	4	84.42	86.28	97.36	98.68
10	5	65.38	80.37	87.74	98.68
10	6	66.41	80.66	87.99	98.29
10	7	64.26	81.59	85.21	98.24
10	8	66.16	81.20	85.11	97.41
10	9	64.60	79.64	85.74	97.80

G.3 Reproduction and compute

The runs use the MPS backend with float32 model arithmetic. The anonymous supplementary package provides the experiment code and dependency versions. From its `experiment` directory, install the dependencies, run the experiment and generate its tables and figure:

```
python -m pip install -r requirements.txt
python scripts/run_grid.py --device mps --out runs
python scripts/summarize_e1_paper.py --runs runs --out results
python scripts/plot_e1_paper.py --summary results/summary.json \
  --out figures/e1_feedback_success.pdf
```

Use a new output directory; `--device cpu` is also supported. The grid uses 800,000 optimizer updates in total. The maximum prediction error is recomputed over all training anchors at checkpoint evaluations. Training generates checkpoints, logs and paired evaluation outcomes locally. Numerical results can vary across compute backends.

Table 3: Selected checkpoints and prediction errors. Losses are averaged over anchors. Every maximum training error exceeds the full-write-back threshold $\min\{\rho, 0.8\}/4$.

H	Seed	Step	Train loss	Val. loss	Max. error	ρ	Threshold
4	0	18500	0.0017	0.0151	2.121	0.606	0.151
4	1	12500	0.0069	0.0043	2.345	0.612	0.153
4	2	16500	0.0019	0.0055	2.082	0.606	0.151
4	3	14500	0.0036	0.0150	2.530	0.593	0.148
4	4	20000	0.0002	0.0082	1.468	0.619	0.155
4	5	19500	0.0007	0.0085	1.963	0.580	0.145
4	6	17500	0.0015	0.0040	1.883	0.599	0.150
4	7	17500	0.0009	0.0095	1.985	0.599	0.150
4	8	18000	0.0017	0.0113	1.959	0.599	0.150
4	9	16000	0.0023	0.0010	2.474	0.619	0.155
6	0	20000	0.0006	0.0201	1.467	0.580	0.145
6	1	20000	0.0004	0.0313	0.800	0.573	0.143
6	2	20000	0.0003	0.0107	0.940	0.573	0.143
6	3	20000	0.0006	0.0168	1.249	0.580	0.145
6	4	20000	0.0006	0.0067	1.773	0.593	0.148
6	5	20000	0.0006	0.0203	0.974	0.580	0.145
6	6	20000	0.0003	0.0311	0.778	0.573	0.143
6	7	20000	0.0007	0.0470	1.251	0.586	0.147
6	8	20000	0.0004	0.0080	0.970	0.586	0.147
6	9	15500	0.0020	0.0287	2.297	0.573	0.143
8	0	20000	0.0064	0.0564	1.807	0.573	0.143
8	1	20000	0.0112	0.0431	1.800	0.573	0.143
8	2	20000	0.0121	0.0603	1.777	0.573	0.143
8	3	20000	0.0085	0.1993	1.801	0.573	0.143
8	4	20000	0.0099	0.0303	1.783	0.573	0.143
8	5	20000	0.0070	0.0368	1.775	0.573	0.143
8	6	20000	0.0088	0.0492	1.792	0.573	0.143
8	7	20000	0.0082	0.0449	1.816	0.573	0.143
8	8	20000	0.0071	0.0444	1.780	0.573	0.143
8	9	20000	0.0062	0.0217	1.876	0.573	0.143
10	0	20000	0.0637	0.1530	1.916	0.573	0.143
10	1	20000	0.0696	0.1856	2.073	0.573	0.143
10	2	20000	0.0566	0.0964	2.418	0.573	0.143
10	3	20000	0.0904	0.1545	1.955	0.573	0.143
10	4	20000	0.0541	0.3327	2.010	0.573	0.143
10	5	20000	0.0921	0.2170	2.031	0.573	0.143
10	6	20000	0.0778	0.1935	2.075	0.573	0.143
10	7	20000	0.0827	0.1637	1.767	0.573	0.143
10	8	20000	0.0675	0.1738	2.169	0.573	0.143
10	9	20000	0.0853	0.1637	1.972	0.573	0.143

Table 4: Full write-back minus raw feedback. Differences and conditional 95% bootstrap intervals are in percentage points; counts include all ten models on the shared environments.

H	Metric	Difference	Interval	Rescued	Lost
4	Path	0.09	$[0.04, 0.15]$	22	3
4	Goal	0.10	$[0.06, 0.15]$	21	1
6	Path	0.08	$[0.03, 0.13]$	17	1
6	Goal	0.07	$[0.02, 0.12]$	18	4
8	Path	0.56	$[0.42, 0.69]$	138	24
8	Goal	0.32	$[0.23, 0.41]$	66	1
10	Path	11.53	$[10.78, 12.28]$	2559	197
10	Goal	8.62	$[8.07, 9.18]$	1853	87

H CORRESPONDENCE BETWEEN MAIN-TEXT STATEMENTS AND APPENDIX RESULTS

Table 5 locates the five main theorems and their proof sources and related results. Setting 1 fixes the search environment; Setting 2 covers execution; Setting 3 covers on-policy learning; Setting 4 covers teacher forcing and full write-back. The appendix gives their full constructions.

Table 5: Proof sources and related results for the five main theorems.

Main theorem / relation	Full statements and proofs
Theorem 1: posterior-moment exponent γ_{obs} and DFS witness (proof source)	Theorem B.6; Definition D.1; Lemmas B.5, B.7, D.2, B.8
Related to Theorem 1	Corollary D.3 (interleaved thoughts)
Theorem 2(a): exact representation, arbitrary valid TI slots (proof source)	Definitions E.1, E.2, E.3; Lemmas E.4, E.5, E.6; Theorem E.8, part (a)
Theorem 2(b): agreement until stopping (proof source)	Theorem E.8(b)
Theorem 3: shared Transformer, finite rounds and $\widetilde{O}(D/N)$ failure (proof source)	Definitions E.7, E.11; Theorem E.8(c) (auxiliary copy-label realizability); Lemmas C.3(ii), E.10; Theorem E.12 (at most K tested rounds; $[\log K + \log(1/\delta_0)]/N$ failure, capped at one)
Theorem 4(a): sufficient condition and ERM (proof source)	Theorem F.2, $\eta_* = \rho/[8S_B(L_{\mathrm{in}})]$, full-column induction at any fixed attention depth, sample-margin derivation and zero ERM training prediction error; Theorem E.8; Lemmas C.2, C.3, C.4 (sharp softmax constant one)
Theorem 4(b): original-interface counterexample (proof source)	Definitions A.4 and F.6 (counterexample class and modification W_a); Lemma F.7 (pair loss and resources); Theorem F.8 (constant initial offset, exact recurrence, exponential near-infimum excess)
Related to Theorem 4	Corollary F.9 (counterexample based on the exact executor)
Theorem 5: robust full write-back, every ERM, sparse H/N and dense H^2/N up to logarithms, and goal reaching (proof source)	Theorem F.2, full-write-back clause including its ERM sentence, with $\rho_{\mathrm{wb}} = \min\{\rho, 4/5\}$; Theorem F.5; Lemma F.1 (uniform restoration under bounded output perturbations); Definition F.3, Corollary F.4; Theorem E.8; Lemmas C.3(iii), C.4, B.8
Relation between Theorems 4(b) and 5	Corollary F.10 (the same near-optimal model and parameters under full write-back, including bounded perturbations)

Assumption 1 supplies the executor hypotheses; Section 5.2 defines the quantized front end. Corollary D.3 extends the single-path bound to interleaved computation. Appendix G documents the experiments: the standard teacher-forcing experiment with raw feedback and full write-back, including all per-seed results.