Internal Tree Search Execution in Transformers

Hibiki Fukushima [1 2] **Taiji Suzuki** [1 2]

Abstract

Tree-structured reasoning can improve language models by exploring multiple intermediate thoughts, but the branching, scoring, and backtracking logic is usually supplied by an external search wrapper. We ask whether this search controller can instead be realized within an autoregressive Transformer itself. We formalize internal tree-search execution as teacher-forced next-token prediction over tokenized search trajectories and study three controlled settings: greedy search on explicit trees, reward-ordered depth-first search on explicit trees, and DFS control over implicit trees generated by a fixed proposal front end. Across these settings, we construct softmax-Transformer controllers that implement the required primitives, including branch comparison, visited-state detection, forward selection, backtracking, and mode routing, under explicit separation and rounding conditions. We further provide finite-sample excess-risk bounds for norm-bounded Transformer classes and show, for explicit trees, that low teacher-forced control risk implies successful rounded autoregressive rollout. These results show that the core control primitives of tree-structured reasoning are representationally and statistically realizable inside Transformer architectures.

1. Introduction

Recent advances in language-model reasoning suggest that performance can improve when models are given additional computation at inference time. Chain-of-thought and scratchpad supervision expose intermediate tokens as a computational workspace (Wei et al., 2022; Nye et al., 2021), while self-consistency and test-time scaling methods improve performance by sampling or allocating multiple reasoning attempts (Wang et al., 2023; Snell et al., 2024). These approaches point to a broader view of reasoning in which useful computation is not limited to extending a single sequence, but may involve exploring several possible intermediate states.

Tree-structured reasoning methods make this exploration explicit. Tree-of-Thought-style algorithms generate candidate thoughts, evaluate partial progress, and search over the resulting tree of reasoning states (Yao et al., 2023). Related planning and graph-based methods similarly combine language models with externally managed search procedures (Hao et al., 2023; Zhou et al., 2024; Besta et al., 2024). In these systems, however, the Transformer typically provides local generative or evaluative steps, while the global controller—the component that manages branching, remembers explored states, selects promising continuations, and backtracks from exhausted branches—is implemented outside the model.

This raises a basic theoretical question: can an autoregressive Transformer execute the control pattern of tree-structured thought search within its own token process? In particular, can it represent partial thoughts, compare candidate continuations, track explored branches, and return to earlier thoughts when the current branch is exhausted? We ask whether the branching, evaluation, and backtracking logic used in Tree-of-Thought-style reasoning can be realized by the model itself, rather than imposed by an external search wrapper.

We study this question by casting tree-structured thought search as teacher-forced next-token prediction over tokenized search trajectories. Explicit trees provide a clean abstraction of branching thought states and let us isolate core control primitives such as branch comparison, visited-state detection, forward selection, and backtracking. We consider three settings: greedy search on explicit trees, reward-ordered DFS on explicit trees, and DFS control over implicit trees whose candidate thoughts and rewards are generated by a fixed front end.

Our results are constructive. For each setting, we build softmax-Transformer controllers that implement the corresponding search transitions under explicit separation and rounding conditions. The constructions use nearly orthogonal sign embeddings for discrete states, so the embedding

[1]Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Hibiki Fukushima <fukushima-hibiki233@g.ecc.u-tokyo.ac.jp>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026.

dimension scales logarithmically with the number of searchable states. We further give finite-sample excess-risk bounds for empirical risk minimization over norm-bounded Transformer classes. Finally, for the explicit-tree settings, we show that low teacher-forced control risk implies successful rounded autoregressive rollout. These results do not claim that realistic language models automatically learn such controllers end-to-end; rather, they show that the core control primitives of tree-structured reasoning are representationally and statistically realizable within Transformer architectures.

1.1. Contributions

A token-level formulation of internal tree search. We formulate tree-structured thought search as teacher-forced next-token prediction over tokenized search trajectories. The formulation separates memory records, selected states, candidate branches, rewards, and transition rules, and covers greedy search, explicit-tree DFS, and implicit-tree DFS within a common autoregressive interface.

Approximation guarantees for softmax-Transformer search controllers. We construct softmax-Transformer controllers that implement the search transitions under explicit separation, temperature, and rounding conditions. The constructions realize reward-based branch comparison, visited-state detection, forward selection, backtracking, and routing, for both explicit trees and implicit trees generated by a fixed front end.

Finite-sample estimation guarantees. We prove excess-risk bounds for empirical risk minimization over norm-bounded Transformer controller classes. These bounds quantify how many teacher-forced search trajectories suffice to learn the constructed transitions statistically, with discrete node identities represented in logarithmic dimension by near-orthogonal sign embeddings.

A bridge from teacher forcing to rounded rollout. For the explicit-tree settings, we show that small teacher-forced control risk implies successful rounded autoregressive execution. This connects the supervised next-token objective to closed-loop search behavior, while keeping clear that our guarantees are constructive rather than claims about unconstrained end-to-end optimization.

1.2. Related work

External search over thoughts. The closest empirical motivation comes from Tree-of-Thought-style reasoning, planning-based language-agent methods, and graph-structured reasoning frameworks (Yao et al., 2023; Hao et al., 2023; Zhou et al., 2024; Besta et al., 2024). These methods use language models as generators, evaluators, or world models inside externally managed search procedures. Our question is different: rather than designing another external search algorithm, we ask whether the search-control primitives used by such methods can be realized by an autoregressive Transformer using its own token sequence as memory.

Learning search from serialized traces. Closest to our motivation on the empirical side is Stream of Search (SoS) (Gandhi et al., 2024), which trains language models on flattened textual search trajectories containing exploration, goal checks, and backtracking. SoS shows that supervising full search traces can outperform supervising only optimal solution paths on Countdown, providing evidence that search behavior can be learned through the language-modeling interface.

Algorithmic Transformers and learned search policies. Prior theoretical work shows that intermediate tokens can increase the computational power of Transformers and support constructive solutions for tasks such as arithmetic, equation solving, dynamic-programming-like computation, and compositional evaluation (Merrill & Sabharwal; Li et al., 2024; Feng et al., 2023; Yehudai et al.). These works mostly study linear traces or fixed-order evaluation, whereas we study adaptive search. A broader literature analyzes algorithmic computation in Transformer-like models (Pérez et al., 2021; Weiss et al., 2021; Lindner et al., 2023; Giannou et al., 2023).

Closest to our explicit-tree DFS setting, De Luca & Fountoulakis (2024) simulate BFS, DFS, Dijkstra, and related graph algorithms using looped Transformers with graph-interacting attention and hardmax-style selection. Their DFS result has a different input–output contract from ours. A graph and its algorithmic state are represented by a node-wise state matrix together with an adjacency matrix; the same Transformer block is repeatedly looped until a termination flag is set; and the final algorithmic output, such as the DFS predecessor array, is decoded from the terminal state. Thus intermediate node choices are represented as internal state updates rather than emitted as a serialized search trace.

Our goal is instead motivated by tree-structured reasoning with language models: each search-control move—forward selection of a child or backtracking from an exhausted node—is itself the next autoregressive token. The evolving search state is stored in the generated token history, not in a recurrent node-wise graph state, and the controller is trained and analyzed through a next-token interface. This changes which overhead is relevant. Looped Transformers share parameters across internal algorithmic iterations and are natural for producing final graph-algorithm outputs on explicit graphs, whereas our construction shares the same controller across autoregressive search tokens but pays for a serialized history. In exchange, our interface is closer to Tree-of-Thought-style generation, reward-ordered tree

search, implicit candidate records, softmax attention, and finite-sample guarantees for learned controllers.

Yang et al. (2025) analyze symbolic multi-step reasoning on trees, while Kim et al. (2026) study unknown-tree search with bandit feedback and externally supplied expansions. Our focus is internal DFS-style search control: the model must compare branches, detect visited states, route between forward moves and backtracking, and, in the implicit setting, control search over generated candidate thoughts. Our results are therefore positive but deliberately controlled: we give constructive softmax implementations and finite-sample risk bounds for teacher-forced internal search execution, rather than claiming that arbitrary search behavior emerges automatically in end-to-end trained language models.

2. Problem Formulation and Model Interface

2.1. Tree instances, rewards, and embeddings

We write a finite tree as $\mathcal{T} = (\mathcal{V}(\mathcal{T}), \mathcal{E}(\mathcal{T}), \mathrm{root}(\mathcal{T}))$, with $|\mathcal{V}(\mathcal{T})| \le S$, where $\mathcal{V}(\mathcal{T})$ is the set of ordinary tree nodes, $\mathcal{E}(\mathcal{T})$ is the set of directed parent-to-child edges, and $\mathrm{root}(\mathcal{T}) \in \mathcal{V}(\mathcal{T})$ is the root. We adjoin a dummy parent node $0 \notin \mathcal{V}(\mathcal{T})$ to the root and write $\mathcal{E}_0(\mathcal{T}) := \mathcal{E}(\mathcal{T}) \cup \{(0, \mathrm{root}(\mathcal{T}))\}$. For $v \in \mathcal{V}(\mathcal{T})$, let $\mathrm{par}(v)$ be the unique node u such that $(u, v) \in \mathcal{E}_0(\mathcal{T})$. For $u \in \{0\} \cup \mathcal{V}(\mathcal{T})$, write $\mathrm{ch}(u) := \{v \in \mathcal{V}(\mathcal{T}) : (u, v) \in \mathcal{E}_0(\mathcal{T})\}$. In the explicit-tree problems, each node $v \in \mathcal{V}(\mathcal{T})$ is assigned a reward $r_{\mathcal{T}}(v) \in \mathbb{R}$, which measures the quality of selecting that node. We assume $|r_{\mathcal{T}}(v)| \le R$. For every node u with $\mathrm{ch}(u) \neq \varnothing$, sibling rewards are separated: $|r_{\mathcal{T}}(v) - r_{\mathcal{T}}(v')| \ge \Delta_{\mathrm{gap}} > 0$ for all $v, v' \in \mathrm{ch}(u)$ with $v \neq v'$. Thus every nonempty subset of siblings has a unique reward maximizer. Each discrete search symbol $q \in \{0, \mathrm{eos}\} \cup \mathcal{V}(\mathcal{T})$ is represented by a sign embedding $a_q \in \{\pm 1/\sqrt{d}\}^d$, for some fixed $\varepsilon \in (0, 1/2)$, satisfying $\langle a_q, a_q \rangle = 1$ and $|\langle a_q, a_{q'} \rangle| \le \varepsilon$ for $q \neq q'$. Such embeddings exist with $d = O(\varepsilon^{-2} \log S)$, and, for trees with branching factor at most $n \ge 2$ and depth at most H, with $d = O(\varepsilon^{-2} H \log n)$; see Appendix A.

2.2. Teacher-forced search-token interfaces

For each problem $j \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$, we specify a token dimension p_j, a memory matrix $M_j(\mathcal{T}) \in \mathbb{R}^{p_j \times m_j(\mathcal{T})}$, and a teacher search-token sequence $o_{j,0}(\mathcal{T}), \ldots, o_{j,L_j(\mathcal{T})}(\mathcal{T}) \in \mathbb{R}^{p_j}$. The initial token $o_{j,0}$ is given in the prefix and is not predicted. At teacher-forced step $k = 0, \ldots, L_j(\mathcal{T}) - 1$, define $E_{j,k}(\mathcal{T}) := [M_j(\mathcal{T}), o_{j,0}(\mathcal{T}), \ldots, o_{j,k}(\mathcal{T})]$; the target is $o_{j,k+1}(\mathcal{T})$. When $M_{\mathrm{III}}(\mathcal{T}) = \emptyset$, $E_{\mathrm{III},k}$ is simply the teacher search-token prefix. Teacher forcing means that $E_{j,k}$ contains the true previous teacher tokens, not the model's previous predictions. We also write $L_{\max,j} := \sup_{\mathcal{T} \in \mathrm{supp}(P_{\mathrm{tr},j})} L_j(\mathcal{T})$, where $P_{\mathrm{tr},j}$ is the training distribution for Problem j.

Problem I: greedy search on an explicit tree. Problem I isolates local reward-based branch selection: the explicit tree and all node rewards are given in the initial memory, and the teacher starts at $u_0 := \mathrm{root}(\mathcal{T})$, repeatedly moves to $u_{k+1} := \arg\max_{v \in \mathrm{ch}(u_k)} r_{\mathcal{T}}(v)$. Let $L_{\mathrm{I}}(\mathcal{T})$ be the number of greedy transitions; since this path never revisits a node, $L_{\max,\mathrm{I}} \le S - 1$. Problem I uses the block layout $x = (P, C, r, m; h_{\mathrm{I}})$, where $P, C \in \mathbb{R}^d$, $r, m \in \mathbb{R}$, and $h_{\mathrm{I}} \in \mathbb{R}^{2d}$; hence $p_{\mathrm{I}} = 4d + 2$. The memory matrix is $M_{\mathrm{I}}(\mathcal{T}) := [(a_u^\top, a_v^\top, r_{\mathcal{T}}(v), 1, 0_d^\top, 0_d^\top)^\top : (u, v) \in \mathcal{E}(\mathcal{T})]$. For $k = 0, \ldots, L_{\mathrm{I}}(\mathcal{T})$, the teacher search-token sequence is $o_{\mathrm{I},k}(\mathcal{T}) := (0_d^\top, a_{u_k}^\top, 0, 0, 0_d^\top, 0_d^\top)^\top$.

Problem II: reward-ordered DFS on an explicit tree. Problem II adds visited-state bookkeeping and backtracking. The explicit tree and all node rewards are given in memory. The DFS teacher path starts from $u_0 := \mathrm{root}(\mathcal{T})$. Given $u_0, \ldots, u_k$, let $\mathsf{V}_k := \{u_0, \ldots, u_k\}$ and $A_k^{\mathrm{fwd}} := \{v \in \mathrm{ch}(u_k) : v \notin \mathsf{V}_k\}$. If $A_k^{\mathrm{fwd}} \neq \emptyset$, the teacher moves to $u_{k+1} := \arg\max_{v \in A_k^{\mathrm{fwd}}} r_{\mathcal{T}}(v)$; if $A_k^{\mathrm{fwd}} = \emptyset$ and $u_k \neq \mathrm{root}(\mathcal{T})$, it backtracks to $u_{k+1} := \mathrm{par}(u_k)$. Let $L_{\mathrm{II}}(\mathcal{T})$ be the number of DFS transitions. Since each tree edge is traversed at most once forward and at most once backward, $L_{\max,\mathrm{II}} \le 2S - 1$. Problem II uses the block layout $x = (P, C, W, r, \chi_{\mathrm{mem}}, \chi_{\mathrm{st}}; h_{\mathrm{II}})$, where $P, C, W \in \mathbb{R}^d$, $r, \chi_{\mathrm{mem}}, \chi_{\mathrm{st}} \in \mathbb{R}$, and $h_{\mathrm{II}} \in \mathbb{R}^{d+3}$; hence $p_{\mathrm{II}} = 4d + 6$. The memory matrix is $M_{\mathrm{II}}(\mathcal{T}) := [(a_u^\top, a_v^\top, 0_d^\top, r_{\mathcal{T}}(v), 1, 0, 0_{d+3}^\top)^\top : (u, v) \in \mathcal{E}_0(\mathcal{T})]$. For $k = 0, \ldots, L_{\mathrm{II}}(\mathcal{T})$, the teacher search-token sequence is $o_{\mathrm{II},k}(\mathcal{T}) := (a_{\mathrm{par}(u_k)}^\top, 0_d^\top, a_{u_k}^\top, 0, 0, 1, 0_{d+3}^\top)^\top$.

A motivation for considering DFS is that there are simple depth-H instances where greedy search fails with high probability, whereas DFS is expected to find the target with $O(H)$ walk cost; see Appendix E.

Problem III: reward-ordered DFS over an implicit generated tree. Problem III has no explicit memory: $M_{\mathrm{III}}(\mathcal{T}) = \emptyset$. Local candidates are produced by a fixed front-end, while the trainable Transformer controls generation, selection, and backtracking. The generated tree has depth at most H, branching factor at most C, and at most $S := \sum_{h=1}^{H} C^{h-1}$ ordinary generated nodes. Let $\mathcal{Z} \subset \mathbb{R}^d$ with $\|z\|_2 \le Z$, and let $e_1, \ldots, e_{C+1}$ be the standard basis of $\mathbb{R}^{C+1}$. Each instance contains an initial thought $z_{\mathrm{root}} \in \mathcal{Z}$ and a root node $\mathrm{root}(\mathcal{T})$. The fixed front-end is a columnwise proposal map $g_{\mathrm{pre}} : \mathcal{Z} \times \mathbb{R}^d \times \{e_1, \ldots, e_{C+1}\} \to \mathcal{Z} \times \mathbb{R}^d \times \mathbb{R}$. For a fixed queried state (z, u), write (z_j, a_{v_j}, r_j) for the j-th proposal, suppressing its dependence on (z, u), and let Valid_j denote the event that $1 \le j \le C$ and this proposal is valid. Then

$$g_{\mathrm{pre}}(z, a_u, e_j) = \begin{cases} (z_j, a_{v_j}, r_j), & \mathsf{Valid}_j, \\ (0_d, a_{\mathrm{eos}}, 0), & \text{otherwise.} \end{cases}$$

Valid thoughts lie in $\mathcal{Z}$, rewards are bounded by R, and valid rewards generated from the same queried state are Δ_{gap}-separated. The teacher operates on selected records (z, π, u, b), where $b = 0$ denotes a fresh visit and $b = 1$ a backtracking return, and candidate records $(z', z_{\mathrm{src}}, \pi, u, v, r, j)$. It starts from $(z_{\mathrm{root}}, 0, \mathrm{root}(\mathcal{T}), 0)$. From a fresh selected record $(z, \pi, u, 0)$, it successively queries $g_{\mathrm{pre}}(z, a_u, e_j)$ for $j = 1, \dots, C$, appending one candidate record per valid proposal until an invalid proposal appears or all C phases are used; backtracking records skip generation. At a DFS-control step, let V_k be the set of current nodes appearing in selected records, and let A_k^{fwd} be the candidate records in the prefix generated from the current node u whose child is not in V_k. If $A_k^{\mathrm{fwd}} \neq \emptyset$, the teacher appends the selected record $(z', u, v, 0)$ for the highest-reward candidate $(z', z, \pi, u, v, r, j) \in A_k^{\mathrm{fwd}}$. If $A_k^{\mathrm{fwd}} = \emptyset$ and $u \neq \mathrm{root}(\mathcal{T})$, then, writing the most recent selected record with current node π as $(z_\pi, \pi', \pi, \cdot)$, the teacher appends $(z_\pi, \pi', \pi, 1)$. Problem III uses the token layout $x = (z, z_{\mathrm{src}}, P, U, V, r, \phi, b, \iota; h_{\mathrm{III}})$, where $z, z_{\mathrm{src}}, P, U, V \in \mathbb{R}^d$, $r, b, \iota \in \mathbb{R}$, $\phi \in \mathbb{R}^{C+1}$, and $h_{\mathrm{III}} \in \mathbb{R}^{7d+5}$; hence $p_{\mathrm{III}} = 12d + C + 9$. Here P, U, V store the parent, current, and candidate-child embeddings, and $\iota = 1$ on legal tokens. Part of h_{III} is reserved for proposal registers (G_z, G_v, G_r). Let G_{pre} be the fixed columnwise lift that, for each token $x = (z, z_{\mathrm{src}}, P, U, V, r, \phi, b, \iota; h_{\mathrm{III}})$, writes $(G_z, G_v, G_r) = g_{\mathrm{pre}}(z_{\mathrm{src}}, U, \phi)$ and leaves the visible blocks unchanged. A selected record is encoded as $s_{\mathrm{sel}}(z, \pi, u, b) := (z^\top, z^\top, a_\pi^\top, a_u^\top, 0_d^\top, 0, e_1^\top, b, 1, 0_{7d+5}^\top)^\top$. The j-th candidate record is encoded as $s_{\mathrm{cand}}(z', z_{\mathrm{src}}, \pi, u, v, r, j) := ((z')^\top, z_{\mathrm{src}}^\top, a_\pi^\top, a_u^\top, a_v^\top, r, e_{j+1}^\top, 0, 1, 0_{7d+5}^\top)^\top$. The initial token is $o_{\mathrm{III},0}(\mathcal{T}) := s_{\mathrm{sel}}(z_{\mathrm{root}}, 0, \mathrm{root}(\mathcal{T}), 0)$. Let $o_{\mathrm{III},0}, \dots, o_{\mathrm{III},L_{\mathrm{III}}}$ be the tokenization of the teacher record sequence. Since each generated node contributes at most one fresh selected visit, at most C candidate tokens, and at most one backtracking return, $L_{\max,\mathrm{III}} \leq (C+2)S - 1$.

2.3. Controller classes and teacher-forced risks

For a token matrix $X = [x_1, \dots, x_m] \in \mathbb{R}^{p \times m}$, and parameters $\mathbf{V} = (V_1, \dots, V_J)$ and $\mathbf{B} = (B_1, \dots, B_J)$, with $V_h, B_h \in \mathbb{R}^{p \times p}$, define a J-head residual attention layer by $\mathcal{A}_{\mathbf{V},\mathbf{B}}(X) := X + \sum_{h=1}^J V_h X \,\mathrm{softmax}(X^\top B_h X)$, where the softmax is applied columnwise. The matrix B_h is the combined key–query score matrix, equivalently $B_h = (K_h)^\top Q_h$. For $\Lambda_V, \Lambda_B > 0$, define

$$\mathfrak{A}_p(J, \Lambda_V, \Lambda_B) := \left\{ \mathcal{A}_{\mathbf{V},\mathbf{B}} \;\middle|\; \begin{array}{l} \|V_h\|_{\mathrm{op}} \leq \Lambda_V,\ \|B_h\|_{\mathrm{op}} \leq \Lambda_B, \\ h = 1, \dots, J \end{array} \right\}.$$

Let $\sigma(t) := \max\{t, 0\}$. For integers $L_{\mathrm{ffn}}, W_{\mathrm{ffn}} \geq 1$, a ReLU FFN $\psi_{\mathbf{A},\mathbf{b}} : \mathbb{R}^p \to \mathbb{R}^p$ is a map $\psi_{\mathbf{A},\mathbf{b}} = (A_{L_{\mathrm{ffn}}} \sigma(\cdot) + b_{L_{\mathrm{ffn}}}) \circ \cdots \circ (A_1 \cdot + b_1)$, where all hidden widths are at most W_{ffn}, and the input and output dimensions are both p. For $\Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}} > 0$, define

$$\Psi_p(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}}) := \left\{ \psi_{\mathbf{A},\mathbf{b}} \;\middle|\; \begin{array}{l} \max_\ell \{\|A_\ell\|_{\mathrm{op}}, \|b_\ell\|_\infty\} \leq \Lambda_{\mathrm{ffn}}, \\ \|\psi_{\mathbf{A},\mathbf{b}}(x)\|_2 \leq c_{\mathrm{ffn}} \|x\|_2 \quad (x \in \mathbb{R}^p) \end{array} \right\}.$$

The corresponding columnwise FFN layer is $\mathcal{F}_\psi(X) := [\psi(x_1), \dots, \psi(x_m)]$.

An architecture budget is denoted by $\mathbf{\Lambda} = (K, J, L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \Lambda_V, \Lambda_B, \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}})$. For problem $j \in \{\mathrm{I}, \mathrm{II}\}$, define the norm-bounded controller class directly by

$$\mathcal{F}_j(\mathbf{\Lambda}_j) := \left\{ \begin{array}{c} \mathcal{F}_{\psi_K} \circ \mathcal{A}_K \circ \cdots \\ \circ \mathcal{F}_{\psi_1} \circ \mathcal{A}_1 \circ \mathcal{F}_{\psi_0} \end{array} \;\middle|\; \begin{array}{c} \mathcal{A}_r \in \mathfrak{A}_{p_j}(J, \Lambda_V, \Lambda_B), \\ \psi_r \in \Psi_{p_j}(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \\ \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}}) \end{array} \right\}.$$

where the components of $\mathbf{\Lambda}_j$ are understood in the displayed order. For Problem III, define

$$\mathcal{F}_{\mathrm{III}}(\mathbf{\Lambda}_{\mathrm{III}}) := \left\{ \begin{array}{c} \mathcal{F}_{\psi_K} \circ \mathcal{A}_K \circ \cdots \\ \circ \mathcal{F}_{\psi_1} \circ \mathcal{A}_1 \circ \mathcal{F}_{\psi_0} \\ \circ G_{\mathrm{pre}} \end{array} \;\middle|\; \begin{array}{c} \mathcal{A}_r \in \mathfrak{A}_{p_{\mathrm{III}}}(J, \Lambda_V, \Lambda_B), \\ \psi_r \in \Psi_{p_{\mathrm{III}}}(L_{\mathrm{ffn}}, W_{\mathrm{ffn}}, \\ \Lambda_{\mathrm{ffn}}, c_{\mathrm{ffn}}) \end{array} \right\}.$$

Here G_{pre} denotes the fixed prefix-level lift defined above; only the attention layers and FFN layers displayed before G_{pre} are trainable and counted in $\mathbf{\Lambda}_{\mathrm{III}}$. For $F \in \mathcal{F}_j(\mathbf{\Lambda}_j)$, define the teacher-forced prediction at step $k+1$ by $\widehat{o}_{F,j,k+1}(\mathcal{T}) := F(E_{j,k}(\mathcal{T}))_{:,-1}$. The teacher-forced population risk is

$$\mathcal{R}_j(F) := \frac{1}{2} \mathbb{E}_{\mathcal{T} \sim P_{\mathrm{tr},j}} \sum_{k=0}^{L_j(\mathcal{T})-1} \|\widehat{o}_{F,j,k+1}(\mathcal{T}) - o_{j,k+1}(\mathcal{T})\|_2^2 .$$

Given i.i.d. samples $\mathcal{T}_1, \dots, \mathcal{T}_N \sim P_{\mathrm{tr},j}$, define

$$\widehat{\mathcal{R}}_{j,N}(F) := \frac{1}{2N} \sum_{i=1}^N \sum_{k=0}^{L_j(\mathcal{T}_i)-1} \|\widehat{o}_{F,j,k+1}(\mathcal{T}_i) - o_{j,k+1}(\mathcal{T}_i)\|_2^2 .$$

An empirical risk minimizer is $\widehat{F}_{j,N} \in \arg\min_{F \in \mathcal{F}_j(\mathbf{\Lambda}_j)} \widehat{\mathcal{R}}_{j,N}(F)$. Let $\mathcal{R}_j^\star$ denote the optimal teacher-forced risk over all measurable predictors. Since the teacher transitions in the three problems are deterministic under the stated tie-free assumptions, $\mathcal{R}_j^\star = 0$. For a fixed budget, define

$$\mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) := \inf_{F \in \mathcal{F}_j(\mathbf{\Lambda}_j)} \mathcal{R}_j(F) - \mathcal{R}_j^\star,$$

$$\mathcal{E}_{\mathrm{est},j,N}(\mathbf{\Lambda}_j) := \mathcal{R}_j(\widehat{F}_{j,N}) - \inf_{F \in \mathcal{F}_j(\mathbf{\Lambda}_j)} \mathcal{R}_j(F).$$

Then

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j^\star = \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \mathcal{E}_{\mathrm{est},j,N}(\mathbf{\Lambda}_j).$$

3. Approximation Analysis: Transformer Constructions

We construct reference controllers for the three teacher-forced transitions defined in Section 2. Construction details are deferred to Appendix B.

For Problem I, the FFNs prepare parent-matching and reward-comparison registers, and the attention head uses them to select the highest-reward outgoing child of the current node. The final FFN formats the selected child as the next selected-state token.

Theorem 3.1 (Exact approximation for Problem I)**.** *Under the Problem I assumptions in Section 2, suppose* $\mathbf{\Lambda}_{\mathrm{I}}$ *satisfies* $K = J = 1$, $L_{\mathrm{ffn}} \geq 2$, $W_{\mathrm{ffn}} \gtrsim d$, $\Lambda_V \gtrsim 1$, $\Lambda_B \gtrsim \log(2S)/((1-2\varepsilon)\Delta_{\mathrm{gap}})$, $\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1+R+\Delta_{\mathrm{gap}}}$, *and* $c_{\mathrm{ffn}} \gtrsim 1 + R + \Delta_{\mathrm{gap}}$. *Then there exists* $F_{\mathrm{I}}^\star \in \mathcal{F}_{\mathrm{I}}(\mathbf{\Lambda}_{\mathrm{I}})$ *such that, for every* $\mathcal{T} \in \mathrm{supp}(P_{\mathrm{tr,I}})$ *and every* $k = 0, \ldots, L_{\mathrm{I}}(\mathcal{T}) - 1$, $F_{\mathrm{I}}^\star(E_{\mathrm{I},k}(\mathcal{T}))_{:,-1} = o_{\mathrm{I},k+1}(\mathcal{T})$. *Hence* $\mathcal{R}_{\mathrm{I}}(F_{\mathrm{I}}^\star) = 0$ *and* $\mathcal{E}_{\mathrm{app,I}}(\mathbf{\Lambda}_{\mathrm{I}}) = 0$.

For Problem II, the first layer separates visited from unvisited child records, the overwrite FFN turns this soft signal into exact candidate registers, and the second layer uses these registers to choose between moving forward to the best unvisited child and backtracking to the appropriate previous state. The final FFN formats the routed state as the next DFS token.

Theorem 3.2 (Exact approximation for Problem II)**.** *Under the Problem II assumptions in Section 2, suppose* $\mathbf{\Lambda}_{\mathrm{II}}$ *satisfies* $K = 2$, $J = 3$, $L_{\mathrm{ffn}} \geq 2$, $W_{\mathrm{ffn}} \gtrsim d$, $\Lambda_V \gtrsim 1$, $\Lambda_B \gtrsim \log(2S)\max\{1, \Delta_{\mathrm{gap}}^{-1}\}/(1-2\varepsilon)$, $\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1+R+\Delta_{\mathrm{gap}}}$, *and* $c_{\mathrm{ffn}} \gtrsim 1 + R + \Delta_{\mathrm{gap}}$. *Then there exists* $F_{\mathrm{II}}^\star \in \mathcal{F}_{\mathrm{II}}(\mathbf{\Lambda}_{\mathrm{II}})$ *such that, for every* $\mathcal{T} \in \mathrm{supp}(P_{\mathrm{tr,II}})$ *and every* $k = 0, \ldots, L_{\mathrm{II}}(\mathcal{T}) - 1$, $F_{\mathrm{II}}^\star(E_{\mathrm{II},k}(\mathcal{T}))_{:,-1} = o_{\mathrm{II},k+1}(\mathcal{T})$. *Hence* $\mathcal{R}_{\mathrm{II}}(F_{\mathrm{II}}^\star) = 0$ *and* $\mathcal{E}_{\mathrm{app,II}}(\mathbf{\Lambda}_{\mathrm{II}}) = 0$.

For Problem III, a fixed front-end first generates candidate thoughts, node embeddings, and rewards from the current source state. The trainable controller then performs selection over these generated candidates and formats the next implicit-tree search token; the fixed front-end is not counted in the trainable budget.

Theorem 3.3 (Approximation for Problem III)**.** *Consider Problem III as defined in Section 2, with the augmented token layout* $p_{\mathrm{III}} = 12d + C + 9$. *Let the generated tree have depth at most* H, *branching factor at most* C, *and* $S := \sum_{h=1}^{H} C^{h-1}$. *Assume that generated thoughts satisfy* $\|z\|_2 \leq Z$, *generated rewards satisfy* $|r| \leq R$, *valid rewards generated from the same source state are* Δ_{gap}*-separated, invalid proposals are represented by* a_{eos}, *and generated node embeddings satisfy the sign-JL condition with* $\varepsilon < 1/2$. *Define* $\gamma_{\mathrm{III}} := \min\{1-\varepsilon, (1-2\varepsilon)\Delta_{\mathrm{gap}}\}$. *Suppose* $\mathbf{\Lambda}_{\mathrm{III}}$ *satisfies* $K = 2$, $J = 3$, $L_{\mathrm{ffn}} \geq 2$, $W_{\mathrm{ffn}} \gtrsim d + C$, $\Lambda_V \gtrsim 1$, $\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1+R+\Delta_{\mathrm{gap}}+Z}$, $c_{\mathrm{ffn}} \gtrsim 1 + R + \Delta_{\mathrm{gap}} + Z$, *and* $\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1}\log((C+2)S)$. *Then there exists* $F_{\mathrm{III}}^\star \in \mathcal{F}_{\mathrm{III}}(\mathbf{\Lambda}_{\mathrm{III}})$ *such that candidate-generation steps are reproduced exactly, and on forward-selection and backtracking steps all discrete node, phase, and flag blocks are rounded exactly. Moreover, with* $L_{\max,\mathrm{III}} \leq (C+2)S - 1$,

$$\mathcal{E}_{\mathrm{app,III}}(\mathbf{\Lambda}_{\mathrm{III}}) \lesssim Z^2 L_{\max,\mathrm{III}} \min\left\{1,\ ((C+2)S)^2 \exp[-\Omega(\gamma_{\mathrm{III}}\Lambda_B)]\right\}.$$

The constructions encode node identities in $d = O(\varepsilon^{-2}\log S)$ dimensions, or $d = O(\varepsilon^{-2} H \log C)$ for depth H and branching factor C. In Problems I and II, softmax attention only has to enter a rounding basin, so the final FFNs recover the discrete search state exactly and the approximation error is zero. In Problem III, discrete control is still rounded exactly, but continuous thought vectors z are transported by attention, leaving the stated softmax-tail approximation error.

The constructive inverse temperatures are not arbitrary isolated parameters. After the FFN preprocessing, value maps, and score directions are fixed as in the proofs, each active head lies in a one-dimensional temperature family $B_h(\lambda_h) = \lambda_h \overline{B}_h$. We tune these temperatures using the source-selection supervision in the teacher-forced trajectories: each active head is trained to put attention mass on the oracle source columns. Appendix B.4 shows that, under the same score-margin conditions used in the approximation proofs, gradient descent on this source-selection loss decreases the loss at rate $O(1/t)$ after a standard warm start. The resulting tail control also bounds transported payloads: if the oracle value is y and all non-oracle values are within distance D of y, then the head-output error is at most D times the non-oracle attention mass; in Problem III, $D \leq 2Z$ for the continuous thought blocks. Thus the large inverse temperatures used above are compatible with elementary gradient-based post-training within the constructed subfamilies.

4. Excess Risk Analysis

We next bound the population excess risk of empirical risk minimization over the norm-bounded controller classes. The approximation results in Section 3 show that the desired search transitions are contained, exactly or up to the stated softmax error, in these classes. This section quantifies the resulting excess risk from finitely many teacher-forced

training trajectories. We state both the standard uniform-convergence bound and the faster Bernstein-type bound; the latter uses the deterministic teacher and the nonnegative bounded squared loss. The proof is deferred to Appendix C.

4.1. Generic excess-risk bound

For a budget $\mathbf{\Lambda}_j = (K_j, J_j, L_{\mathrm{ffn},j}, W_{\mathrm{ffn},j}, \Lambda_{V,j}, \Lambda_{B,j}, \Lambda_{\mathrm{ffn},j}, c_{\mathrm{ffn},j})$, let $D_j = D_j(\mathbf{\Lambda}_j)$ be the number of trainable scalar parameters. For the controller template in Section 2.3, one may take $D_j = O\Big(K_j J_j p_j^2 + (K_j+1)L_{\mathrm{ffn},j}(W_{\mathrm{ffn},j}^2 + p_j W_{\mathrm{ffn},j} + p_j)\Big)$. For Problem III, the fixed local proposal lift G_{pre}, equivalently the columnwise operation $(G_z, G_v, G_r) = g_{\mathrm{pre}}(z_{\mathrm{src}}, U, \phi)$, is not counted in D_{III}.

Let $\mathsf{B}_j(\mathbf{\Lambda}_j)$ be a uniform envelope for the per-instance teacher-forced loss, i.e., $\ell_{F,j}(\mathcal{T}) := \frac{1}{2}\sum_{k=0}^{L_j(\mathcal{T})-1} \|F(E_{j,k}(\mathcal{T}))_{:,-1} - o_{j,k+1}(\mathcal{T})\|_2^2 \le \mathsf{B}_j$ for all $F \in \mathcal{F}_j(\mathbf{\Lambda}_j)$ and all $\mathcal{T}$ in the support. Under the norm constraints, it is enough to take $\mathsf{B}_j(\mathbf{\Lambda}_j) := \frac{L_{\max,j}}{2}\left(c_{\mathrm{ffn},j}^{K_j+1}(1+J_j\Lambda_{V,j})^{K_j}B_{x,j} + B_{y,j}\right)^2$, where $B_{x,j}$ bounds the column norm of the input seen by the trainable controller and $B_{y,j}$ bounds the target-token norm. In Problem III, $B_{x,\mathrm{III}}$ is taken after the fixed lift G_{pre}. In our three settings, we may take $(B_{x,\mathrm{I}}, B_{y,\mathrm{I}}) = (\sqrt{3+R^2}, 1)$, $(B_{x,\mathrm{II}}, B_{y,\mathrm{II}}) = (\sqrt{3+R^2}, \sqrt{3})$, and $B_{x,\mathrm{III}} = O(1+Z+R)$, $B_{y,\mathrm{III}} = \sqrt{2Z^2+R^2+5}$.

Theorem 4.1 (Excess-risk bounds for norm-bounded search controllers)**.** *Fix $j \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ and a budget $\mathbf{\Lambda}_j$. Let $\widehat{F}_{j,N} \in \arg\min_{F\in\mathcal{F}_j(\mathbf{\Lambda}_j)} \widehat{\mathcal{R}}_{j,N}(F)$. Then, with probability at least $1-\delta$,*

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j^\star \le \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\sqrt{\frac{D_j(\mathbf{\Lambda}_j)+\log(1/\delta)}{N}}\right).$$

Moreover, since the teacher transitions are deterministic and $\mathcal{R}_j^\star = 0$, the same ERM satisfies the Bernstein-type excess-risk bound

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j^\star \le 4\mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\frac{D_j(\mathbf{\Lambda}_j)+\log(1/\delta)}{N}\right).$$

Here $\widetilde{O}(\cdot)$ hides only logarithmic factors in $N, \delta^{-1}, p_j, L_{\max,j}$, the token radii, and the norm radii in $\mathbf{\Lambda}_j$.

The two displays give slow- and fast-rate upper bounds on the same population excess risk. The first follows from uniform convergence, while the second uses the Bernstein condition for bounded nonnegative losses with deterministic targets. In the problem-level rates below, we use the sharper of the two bounds.

4.2. Problem-level excess-risk rates

Let $d_S := \varepsilon^{-2}\log(S+2)$ and $\ell_\delta := \log(1/\delta)$. We use the approximation-feasible budgets from Theorems 3.1, 3.2, and 3.3, taken at the smallest required orders. Thus $d \asymp d_S$ for the near-orthogonal node dictionary, and the required FFN widths satisfy $W_{\mathrm{ffn,I}}, W_{\mathrm{ffn,II}} \asymp d_S$ and $W_{\mathrm{ffn,III}} \asymp C + d_S$. Set $M_{\mathrm{ex}} := 1+R+\Delta_{\mathrm{gap}}$ and $M_{\mathrm{imp}} := 1+R+\Delta_{\mathrm{gap}}+Z$.

Problem I. For the greedy-search budget of Theorem 3.1, the approximation error is zero. Hence, with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{I}}(\widehat{F}_{\mathrm{I},N}) - \mathcal{R}_{\mathrm{I}}^\star = \widetilde{O}\left(SM_{\mathrm{ex}}^6 \min\left\{\frac{d_S+\sqrt{\ell_\delta}}{\sqrt{N}}, \frac{d_S^2+\ell_\delta}{N}\right\}\right).$$

Problem II. For the DFS budget of Theorem 3.2, the approximation error is also zero. Thus, with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{II}}(\widehat{F}_{\mathrm{II},N}) - \mathcal{R}_{\mathrm{II}}^\star = \widetilde{O}\left(SM_{\mathrm{ex}}^8 \min\left\{\frac{d_S+\sqrt{\ell_\delta}}{\sqrt{N}}, \frac{d_S^2+\ell_\delta}{N}\right\}\right).$$

Problem III. For Problem III, the generated tree has depth at most H, branching factor at most C, and $S = \sum_{h=1}^H C^{h-1}$. Let $d_{\mathrm{imp}} := C + d_S$. For a fixed attention-score radius Λ_B, write $\mathcal{E}_{\mathrm{app,III}}(\Lambda_B)$ for the approximation term. Theorems 4.1 and 3.3 give, with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{III}}(\widehat{F}_{\mathrm{III},N}) - \mathcal{R}_{\mathrm{III}}^\star \le \min\begin{Bmatrix}\mathcal{E}_{\mathrm{app,III}}(\Lambda_B) + \mathfrak{e}_{\sqrt{N}}(\Lambda_B), \\ 4\mathcal{E}_{\mathrm{app,III}}(\Lambda_B) + \mathfrak{e}_N(\Lambda_B)\end{Bmatrix},$$

where

$$\mathfrak{e}_{\sqrt{N}}(\Lambda_B) = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8\frac{d_{\mathrm{imp}}+\sqrt{\ell_\delta}}{\sqrt{N}}\right),$$

$$\mathfrak{e}_N(\Lambda_B) = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8\frac{d_{\mathrm{imp}}^2+\ell_\delta}{N}\right).$$

The hidden factors in these two terms include only logarithmic dependence on Λ_B.

It remains to choose Λ_B so that the approximation term is absorbed into the statistical scale. Define $\gamma_{\mathrm{III}} := \min\{1-\varepsilon, (1-2\varepsilon)\Delta_{\mathrm{gap}}\}$ and

$$\bar{\mathfrak{e}}_{\mathrm{III}} := (C+2)SM_{\mathrm{imp}}^8 \min\left\{\frac{d_{\mathrm{imp}}+\sqrt{\ell_\delta}}{\sqrt{N}}, \frac{d_{\mathrm{imp}}^2+\ell_\delta}{N}\right\}.$$

Using $L_{\max,\mathrm{III}} \le (C+2)S-1$, it is sufficient to take

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1} \max\left\{ \begin{array}{l} \log((C+2)S), \\ \log\left(1+\dfrac{Z^2((C+2)S)^3}{\bar{\mathfrak{e}}_{\mathrm{III}}}\right) \end{array} \right\}.$$

With this choice, the approximation term is absorbed into the statistical term, and with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{III}}(\widehat{F}_{\mathrm{III},N}) - \mathcal{R}^\star_{\mathrm{III}} = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8 \min\left\{ \begin{array}{l} \dfrac{C+d_S+\sqrt{\ell_\delta}}{\sqrt{N}}, \\ \dfrac{(C+d_S)^2+\ell_\delta}{N} \end{array} \right\} \right).$$

The hidden logarithmic factors include the logarithmic dependence on the balanced Λ_B.

For Problems I and II, the feasible budgets have zero approximation error, so the excess risk is purely statistical. Up to reward- and norm-dependent constants, Problem II scales as $\widetilde{O}(S\min\{d_S/\sqrt{N}, d_S^2/N\})$, where the factor S comes from the DFS trajectory length and $d_S = O(\varepsilon^{-2}\log S)$ from the node dictionary. Thus the representation of node identities is logarithmic in the search space, while the supervised loss still sums over the search trajectory. For Problem III, choosing Λ_B logarithmically large absorbs the continuous-payload approximation error into the estimation scale. These are teacher-forced excess-risk guarantees, not end-to-end optimization guarantees.

5. Rounded Autoregressive Execution

The excess-risk bounds above are teacher-forced. To relate them to closed-loop execution, we consider Problems I and II with external rounding of the discrete control blocks after each autoregressive prediction.

Rounding and control risk. Let $\Pi_{\mathcal{T}}(z) \in \arg\min_{q\in\{0,\mathrm{eos}\}\cup\mathcal{V}(\mathcal{T})} \|z - a_q\|_2$ be nearest-neighbor rounding in the node dictionary. Since $\|a_q - a_{q'}\|_2^2 \ge 2(1-\varepsilon)$ for $q \ne q'$, the correct node is recovered whenever $\|z-a_q\|_2^2 < (1-\varepsilon)/2$. Let $\mathsf{Q}_{j,\mathcal{T},k+1}$ be the legal-token projector that rounds active node blocks by $\Pi_{\mathcal{T}}$, rounds binary blocks to $\{0,1\}$, and resets inactive blocks and scratch registers to their canonical values. The corresponding rounding radius is $\delta^2_{\mathrm{rd},j} := \min\{(1-\varepsilon)/2, 1/4\}$, $j \in \{\mathrm{I},\mathrm{II}\}$. Under $\varepsilon < 1/2$, this is $\delta^2_{\mathrm{rd},j} = 1/4$.

Let $\mathsf{D}_{j,k,\mathcal{T}}$ extract the discrete control blocks that must be rounded correctly at step $k+1$. Define

$$\mathcal{R}_j^{\mathrm{ctrl}}(F) := \frac{1}{2}\mathbb{E}_{\mathcal{T}\sim P_{\mathrm{tr},j}} \sum_{k=0}^{L_j(\mathcal{T})-1} \left\| \begin{array}{r} \mathsf{D}_{j,k,\mathcal{T}}(F(E_{j,k}(\mathcal{T}))_{:,-1}) \\ -\,\mathsf{D}_{j,k,\mathcal{T}}(o_{j,k+1}(\mathcal{T})) \end{array} \right\|_2^2.$$

Then $\mathcal{R}_j^{\mathrm{ctrl}}(F) \le \mathcal{R}_j(F)$. Starting from $\widehat{o}_{j,0}(\mathcal{T}) = o_{j,0}(\mathcal{T})$, define the rounded rollout recursively by

$$\widehat{o}_{j,k+1}(\mathcal{T}) = \mathsf{Q}_{j,\mathcal{T},k+1}\Big(F(\widehat{E}_{j,k}(\mathcal{T}))_{:,-1}\Big),$$

where $\widehat{E}_{j,k}$ is the prefix formed from the memory and the previously rounded tokens. Let $\mathsf{Succ}_j^{\mathrm{ctrl}}(F,\mathcal{T})$ be the event that the rounded rollout matches the teacher trajectory on all discrete control blocks.

Theorem 5.1 (Teacher-forced control risk controls rounded rollout)**.** *Fix $j \in \{\mathrm{I},\mathrm{II}\}$ and $F \in \mathcal{F}_j(\mathbf{\Lambda}_j)$. Then*

$$\mathbb{P}_{\mathcal{T}\sim P_{\mathrm{tr},j}}\Big(\mathsf{Succ}_j^{\mathrm{ctrl}}(F,\mathcal{T})\Big) \ge 1 - \frac{2\mathcal{R}_j^{\mathrm{ctrl}}(F)}{\delta^2_{\mathrm{rd},j}}.$$

In particular, the same bound holds with $\mathcal{R}_j(F)$ in place of $\mathcal{R}_j^{\mathrm{ctrl}}(F)$.

The proof is given in Appendix D. Combining the theorem with the excess-risk rates from Section 4 gives the following ERM rollout guarantee.

Corollary 5.2 (ERM-to-rounded-rollout guarantees)**.** *Use the approximation-feasible budgets from Section 4. Let $d_S := \varepsilon^{-2}\log(S+2)$, $\ell_\eta := \log(1/\eta)$, and $M_{\mathrm{ex}} := 1 + R + \Delta_{\mathrm{gap}}$. Define*

$$\mathfrak{r}_{\mathrm{I}} := SM_{\mathrm{ex}}^6 \min\left\{ \frac{d_S+\sqrt{\ell_\eta}}{\sqrt{N}}, \frac{d_S^2+\ell_\eta}{N} \right\},$$

$$\mathfrak{r}_{\mathrm{II}} := SM_{\mathrm{ex}}^8 \min\left\{ \frac{d_S+\sqrt{\ell_\eta}}{\sqrt{N}}, \frac{d_S^2+\ell_\eta}{N} \right\}.$$

Then, with probability at least $1-\eta$ over the training sample, for a fresh test instance $\mathcal{T}$,

$$\mathbb{P}_{\mathcal{T}}\Big(\mathsf{Succ}_j^{\mathrm{ctrl}}(\widehat{F}_{j,N},\mathcal{T})\Big) \ge 1 - \widetilde{O}\left(\frac{\mathfrak{r}_j}{\delta^2_{\mathrm{rd},j}}\right), \quad j \in \{\mathrm{I},\mathrm{II}\}.$$

Since $\delta^2_{\mathrm{rd},j} = 1/4$ under the standing rounding convention, this is $1 - \widetilde{O}(\mathfrak{r}_j)$. The hidden logarithmic factors are the same as in Section 4.

Remark on Problem III. Problem III is excluded from this rounded-rollout statement. There, the continuous blocks z, z_{src} and r are fed back into candidate generation and reward-based selection, so correct discrete rounding alone does not determine the future teacher trajectory. A closed-loop statement relative to the model-generated candidates and rewards would require a separate on-policy formulation.

6. Numerical Experiments

We conduct two controlled experiments for Problem II. Both experiments use synthetic reward-labeled complete trees and evaluate teacher-forced next-token prediction. They are intended as diagnostics for the constructive and supervised-learning claims of the theory, rather than as experiments on external language data, autoregressive rollout, or out-of-distribution transfer.

First, we evaluate the explicitly constructed DFS controller while varying the common attention-score radius Λ_B. We use a fixed complete 4-ary tree of depth 4, with fixed node embeddings, and average over 30 independently sampled reward assignments. For each parent node, the rewards of its C children are given by a random permutation of

$$\{0, \Delta_{\mathrm{gap}}, 2\Delta_{\mathrm{gap}}, \ldots, (C-1)\Delta_{\mathrm{gap}}\},$$

with $\Delta_{\mathrm{gap}} = 1$. For each Λ_B, we measure the largest teacher-forced prediction error before rounding, averaged over the 30 samples, and the fraction of samples for which rounding recovers the entire teacher DFS trajectory.

Second, we train a bilinear-attention controller from random initialization and evaluate its teacher-forced test performance as the number of training trees increases. In this experiment, $C = 3$, $H = 4$, so the tree has $S = 40$ nodes. The training and test sets use the same complete-tree topology and the same node embeddings, but have independently sampled sibling reward permutations. Thus, this is a held-out test from the same synthetic data-generating distribution as the training set; it does not test generalization to new topologies or new embeddings.

For each reward-labeled tree, the target sequence is the reward-ordered DFS trajectory from Problem II. Since $S = 40$, each tree gives a trajectory of length $2S - 1 = 79$, hence 78 supervised next-token prediction examples. At each step, the input consists of the explicit edge tokens describing the tree and the previous tokens from the teacher DFS trajectory, and the target is the next token in that trajectory. The trainable parameters are those of the bilinear-attention controller, including the attention matrices, value matrices, feed-forward layers, layer-normalization parameters, and final output map; the tree, rewards, node embeddings, and teacher trajectories are fixed data.

Figure 1 shows the expected behavior. For the explicitly constructed controller, the mean prediction error before rounding falls below the rounding threshold as Λ_B increases, after which the rounded predictions recover the full teacher DFS trajectory on the sampled trees. For the learned controller, increasing the number of training trees from 16 to 1024 reduces the test mean-squared prediction error from 0.077 to 0.0038, and the accuracy of simultaneously predicting the parent and current nodes increases from 0.60 to 1.00.

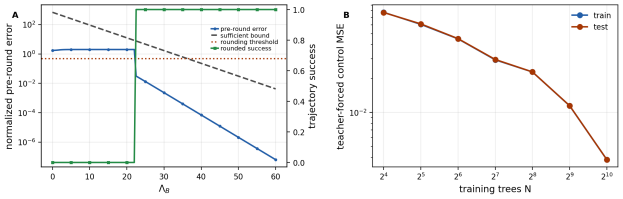

Figure 1. Problem II experiments. Left: evaluation of the explicitly constructed DFS controller on a fixed complete 4-ary tree of depth 4, averaged over 30 independently sampled reward assignments. The blue curve shows the mean prediction error before rounding, the dotted line is the rounding threshold $1/2$, the black dashed curve is the sufficient softmax-tail bound, and the green curve shows the fraction of samples for which the rounded predictions match the entire teacher DFS trajectory. Right: teacher-forced learning curve for a trainable bilinear-attention controller on complete 3-ary trees of depth 4. Training and test trees share the same topology and node embeddings and differ only in independently sampled sibling reward permutations. The reported learning curve uses one optimization seed.

These results support the teacher-forced supervised formulation in this controlled Problem II setting, but they do not constitute an end-to-end optimization guarantee or an out-of-distribution generalization result.

7. Conclusion

We studied whether the search-control logic underlying tree-structured reasoning can be executed inside an autoregressive Transformer. By formulating tree search as teacher-forced next-token prediction over tokenized search trajectories, we isolated the basic operations required for internal search execution: comparing candidate branches, detecting visited states, selecting forward moves, backtracking, and routing between search modes.

For each problem, we constructed softmax-Transformer controllers that realize these operations under explicit separation and rounding conditions. We further established finite-sample excess-risk bounds for norm-bounded controller classes, and showed that small teacher-forced control risk implies successful rounded autoregressive rollout in the explicit-tree settings.

These results indicate that the algorithmic components of Tree-of-Thought-style search are not necessarily tied to an external search wrapper: in controlled settings, they can be represented and statistically learned within Transformer architectures. Our analysis is constructive and teacher-forced, and does not claim that such controllers are automatically found by end-to-end training in realistic language-model systems. Extending these guarantees to on-policy execution, learned proposal models, and less idealized reasoning environments remains an important direction for future work.

References

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

De Luca, A. B. and Fountoulakis, K. Simulation of graph algorithms with looped transformers. In *International Conference on Machine Learning*, pp. 2319–2363. PMLR, 2024.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

Gandhi, K., Lee, D. H. J., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. Stream of search (sos): Learning to search in language. In *First Conference on Language Modeling*, 2024.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D., and Hu, Z. 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.

Kim, J., Zeng, T., Lin, Z., Lee, M., Lee, C., Sohn, J.-y., Koo, H. I., and Lee, K. Transformers in the dark: Navigating unknown search spaces via bandit feedback. *arXiv preprint arXiv:2603.24780*, 2026.

Li, Z., Liu, H., Zhou, D., and Ma, T. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Lindner, D., Kramár, J., Farquhar, S., Rahtz, M., McGrath, T., and Mikulik, V. Tracr: Compiled transformers as a laboratory for interpretability. *Advances in Neural Information Processing Systems*, 36:37876–37899, 2023.

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

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. Show your work: Scratchpads for intermediate computation with language models. 2021.

Pérez, J., Barceló, P., and Marinkovic, J. Attention is turing-complete. *Journal of Machine Learning Research*, 22 (75):1–35, 2021.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Weiss, G., Goldberg, Y., and Yahav, E. Thinking like transformers. In *International Conference on Machine Learning*, pp. 11080–11090. PMLR, 2021.

Yang, T., Huang, Y., Liang, Y., and Chi, Y. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. *arXiv preprint arXiv:2508.08222*, 2025.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Yehudai, G., Amsel, N., and Bruna, J. Compositional reasoning with transformers, rnns, and chain of thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. 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.

A. Rademacher Sign Dictionaries

This appendix proves the existence of the near-orthogonal sign embeddings used in Section 2.1. The argument is a standard Rademacher construction based on concentration and a union bound.

Lemma A.1 (Search-state Rademacher dictionary)**.** *Let $\mathfrak{T}_S$ be a family of search problems with at most S ordinary nodes, and for each instance $\mathcal{T} \in \mathfrak{T}_S$, write*

$$\mathcal{I}_{\mathrm{emb}}(\mathcal{T}) := \{0, \mathrm{eos}\} \cup \mathcal{V}(\mathcal{T}).$$

Fix $\varepsilon \in (0,1)$. For every instance $\mathcal{T} \in \mathfrak{T}_S$, there exist embeddings

$$a_q \in \left\{ \pm \frac{1}{\sqrt{d}} \right\}^d, \qquad q \in \mathcal{I}_{\mathrm{emb}}(\mathcal{T}),$$

such that

$$\langle a_q, a_q \rangle = 1 \qquad (q \in \mathcal{I}_{\mathrm{emb}}(\mathcal{T})),$$

and

$$|\langle a_q, a_{q'} \rangle| \le \varepsilon \qquad (q \neq q').$$

This can be achieved with

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

In particular,

$$d = O\big(\varepsilon^{-2} \log(S+2)\big),$$

and, for $S \ge 2$,

$$d = O\big(\varepsilon^{-2} \log S\big).$$

Moreover, in the explicit-tree setting, if the branching factor is at most $n \ge 2$ and the tree has at most $H \ge 1$ levels, so that the number of ordinary nodes is bounded by $1 + n + \cdots + n^{H-1}$, then the embeddings can be chosen with

$$d = O\big(\varepsilon^{-2} H \log n\big).$$

Proof. Fix an instance $\mathcal{T} \in \mathfrak{T}_S$. Since

$$\mathcal{I}_{\mathrm{emb}}(\mathcal{T}) = \{0, \mathrm{eos}\} \cup \mathcal{V}(\mathcal{T}),$$

we have

$$|\mathcal{I}_{\mathrm{emb}}(\mathcal{T})| \le S + 2.$$

For each symbol $q \in \mathcal{I}_{\mathrm{emb}}(\mathcal{T})$ and each coordinate $\ell \in [d]$, draw independent Rademacher random variables

$$\sigma_{q,\ell} \in \{-1, +1\}, \qquad \mathbb{P}(\sigma_{q,\ell} = 1) = \mathbb{P}(\sigma_{q,\ell} = -1) = \frac{1}{2},$$

and set

$$a_q := \frac{1}{\sqrt{d}} (\sigma_{q,1}, \ldots, \sigma_{q,d})^\top.$$

Then

$$a_q \in \left\{ \pm \frac{1}{\sqrt{d}} \right\}^d$$

and

$$\langle a_q, a_q \rangle = \frac{1}{d} \sum_{\ell=1}^{d} \sigma_{q,\ell}^2 = 1.$$

Now fix two distinct symbols $q \neq q'$. Their inner product is

$$\langle a_q, a_{q'} \rangle = \frac{1}{d} \sum_{\ell=1}^{d} \sigma_{q,\ell} \sigma_{q',\ell}.$$

The variables

$$X_\ell := \sigma_{q,\ell}\sigma_{q',\ell}$$

are independent Rademacher random variables. Hence, by Hoeffding's inequality,

$$\mathbb{P}\left(|\langle a_q, a_{q'}\rangle| > \varepsilon\right) = \mathbb{P}\left(\left|\sum_{\ell=1}^{d} X_\ell\right| > d\varepsilon\right) \le 2\exp\left(-\frac{d\varepsilon^2}{2}\right).$$

Taking a union bound over all unordered pairs in $\mathcal{I}_{\mathrm{emb}}(\mathcal{T})$, we obtain

$$\mathbb{P}\left(\exists q \neq q' : |\langle a_q, a_{q'}\rangle| > \varepsilon\right) \le |\mathcal{I}_{\mathrm{emb}}(\mathcal{T})|\big(|\mathcal{I}_{\mathrm{emb}}(\mathcal{T})| - 1\big)\exp\left(-\frac{d\varepsilon^2}{2}\right).$$

Since

$$|\mathcal{I}_{\mathrm{emb}}(\mathcal{T})| \le S + 2,$$

the right-hand side is bounded by

$$(S+2)(S+1)\exp\left(-\frac{d\varepsilon^2}{2}\right).$$

With the stated choice of d, this is at most $1/2$. Therefore, with positive probability, all pairwise inner products are bounded by ε. Hence at least one deterministic realization of the sampled vectors satisfies all desired properties.

It remains only to check the branching-depth specialization. If every explicit tree has branching factor at most $n \ge 2$ and at most $H \ge 1$ levels, then the number of ordinary nodes is bounded by

$$1 + n + \cdots + n^{H-1} = \frac{n^H - 1}{n - 1}.$$

Thus

$$|\mathcal{I}_{\mathrm{emb}}(\mathcal{T})| \le \frac{n^H - 1}{n - 1} + 2.$$

Applying the preceding construction with this bound in place of $S + 2$ gives

$$d = O\left(\varepsilon^{-2}\log\left(\frac{n^H - 1}{n - 1} + 2\right)\right).$$

Since $n \ge 2$ and $H \ge 1$,

$$\log\left(\frac{n^H - 1}{n - 1} + 2\right) = O(H \log n),$$

and therefore

$$d = O\big(\varepsilon^{-2} H \log n\big).$$

□

B. Approximation Analysis: Full Constructions

B.1. Problem I: greedy search on an explicit tree

This subsection proves Theorem 3.1. The proof is for arbitrary explicit trees with at most S ordinary nodes. The specialization to bounded branching and depth follows by substituting $S \le S_{n,H} = 1 + n + \cdots + n^{H-1}$.

We construct an explicit FFN–attention–FFN controller. The first FFN is a direct preprocessing map, not a residual increment. This is important because the FFN layers in $\mathcal{F}_{\mathrm{I}}(\mathbf{\Lambda}_{\mathrm{I}})$ are columnwise maps without FFN skip connections, while the attention layer itself is residual.

Block notation. Throughout this subsection, write

$$p = p_{\mathrm{I}} = 4d + 2$$

and decompose

$$\mathbb{R}^p = \underbrace{\mathbb{R}^d}_{x_1} \oplus \underbrace{\mathbb{R}^d}_{x_2} \oplus \underbrace{\mathbb{R}}_{x_3} \oplus \underbrace{\mathbb{R}}_{x_4} \oplus \underbrace{\mathbb{R}^d}_{x_5} \oplus \underbrace{\mathbb{R}^d}_{x_6}.$$

For

$$x = (x_1, x_2, x_3, x_4, x_5, x_6)^\top \in \mathbb{R}^p,$$

the blocks have the following roles:

$$x_1 = \text{parent embedding}, \qquad x_2 = \text{child/current embedding}, \qquad x_3 = \text{reward register},$$

$$x_4 = \text{memory/query indicator}, \qquad x_5 = \text{reward-weighted parent register}, \qquad x_6 = \text{attention summary register}.$$

Internal construction constants. The main theorem only states the order of the required budgets. In this proof we fix the following construction-specific constants:

$$\kappa_\varepsilon := 1 - 2\varepsilon, \qquad M_{\mathrm{I}} := 1 + R + \Delta_{\mathrm{gap}},$$

$$\tau_{\mathrm{I}} := R + \Delta_{\mathrm{gap}}, \qquad \lambda_{\mathrm{I}} := 2R + \Delta_{\mathrm{gap}} + 1, \qquad \mu_{\mathrm{I}} := \frac{1}{2}.$$

Then

$$\tau_{\mathrm{I}} - R = \Delta_{\mathrm{gap}}, \qquad \lambda_{\mathrm{I}} > \tau_{\mathrm{I}} + R,$$

and

$$\tau_{\mathrm{I}} \le M_{\mathrm{I}}, \qquad \lambda_{\mathrm{I}} \le 2M_{\mathrm{I}}.$$

These auxiliary constants are used only in the proof.

The preprocessing FFN. Define

$$a_\tau(x) := x_3 + \tau_{\mathrm{I}} x_4, \qquad s_\ell(x) := \sqrt{d}\, x_{1,\ell} \quad (\ell = 1, \dots, d).$$

Define the state-token detector

$$C_{\mathrm{st}}(x) := \frac{1}{d} \sum_{\ell=1}^{d} \left[\sigma(\sqrt{d}\, x_{2,\ell} - x_4) + \sigma(-\sqrt{d}\, x_{2,\ell} - x_4) \right].$$

For each $\ell = 1, \dots, d$, define

$$P_\ell(x) := \frac{1}{\sqrt{d}} \left[\sigma(a_\tau(x) + \lambda_{\mathrm{I}} s_\ell(x)) - \sigma(\lambda_{\mathrm{I}} s_\ell(x)) - \sigma(a_\tau(x) - \lambda_{\mathrm{I}} s_\ell(x)) + \sigma(-\lambda_{\mathrm{I}} s_\ell(x)) \right],$$

and let

$$P(x) := (P_1(x), \dots, P_d(x))^\top.$$

The first FFN is the columnwise lift of

$$\psi_{\mathrm{prep}}(x) := \big(x_1,\ x_2,\ x_3 + \tau_{\mathrm{I}} x_4,\ x_4 + C_{\mathrm{st}}(x),\ P(x),\ 0_d\big).$$

Lemma B.1 (Exact preprocessing on Problem I tokens)**.** *For every memory token*

$$m_{\mathrm{I}}(u, v) = (a_u, a_v, r_{\mathcal{T}}(v), 1, 0_d, 0_d),$$

one has

$$\psi_{\mathrm{prep}}\big(m_{\mathrm{I}}(u, v)\big) = \big(a_u,\ a_v,\ r_{\mathcal{T}}(v) + \tau_{\mathrm{I}},\ 1,\ (r_{\mathcal{T}}(v) + \tau_{\mathrm{I}}) a_u,\ 0_d\big).$$

For every selected-state token

$$s_{\mathrm{I}}(u) = (0_d, a_u, 0, 0, 0_d, 0_d),$$

one has

$$\psi_{\mathrm{prep}}\big(s_{\mathrm{I}}(u)\big) = (0_d, a_u, 0, 1, 0_d, 0_d).$$

Proof. First consider a memory token $x = m_{\mathrm{I}}(u, v)$. Since $x_2 = a_v \in \{\pm 1/\sqrt{d}\}^d$ and $x_4 = 1$, for each coordinate

$$\sigma(\sqrt{d}\, x_{2,\ell} - 1) + \sigma(-\sqrt{d}\, x_{2,\ell} - 1) = 0.$$

Thus $C_{\mathrm{st}}(x) = 0$, and the fourth block remains equal to 1.

Moreover,

$$a_\tau(x) = r_{\mathcal{T}}(v) + \tau_{\mathrm{I}}, \qquad s_\ell(x) = \sqrt{d}\,(a_u)_\ell \in \{+1, -1\}.$$

Since $|r_{\mathcal{T}}(v)| \le R$,

$$|a_\tau(x)| \le R + \tau_{\mathrm{I}} < \lambda_{\mathrm{I}}.$$

If $s_\ell(x) = 1$, then

$$P_\ell(x) = \frac{1}{\sqrt{d}}\left[(a_\tau(x) + \lambda_{\mathrm{I}}) - \lambda_{\mathrm{I}} - 0 + 0\right] = \frac{a_\tau(x)}{\sqrt{d}}.$$

If $s_\ell(x) = -1$, then

$$P_\ell(x) = \frac{1}{\sqrt{d}}\left[0 - 0 - (a_\tau(x) + \lambda_{\mathrm{I}}) + \lambda_{\mathrm{I}}\right] = -\frac{a_\tau(x)}{\sqrt{d}}.$$

Hence

$$P_\ell(x) = a_\tau(x)(a_u)_\ell,$$

and therefore

$$P(x) = (r_{\mathcal{T}}(v) + \tau_{\mathrm{I}})a_u.$$

Now consider a selected-state token $x = s_{\mathrm{I}}(u)$. Then $x_4 = 0$ and $x_2 = a_u$, so for each coordinate

$$\sigma(\sqrt{d}\, x_{2,\ell}) + \sigma(-\sqrt{d}\, x_{2,\ell}) = 1.$$

Thus $C_{\mathrm{st}}(x) = 1$. Also $x_1 = 0_d$, $x_3 = 0$, and $x_4 = 0$, so $a_\tau(x) = 0$ and $s_\ell(x) = 0$. Hence $P_\ell(x) = 0$ for all ℓ. This proves the claim. □

The selection attention head. Set

$$C_{\mathrm{I}} := 2S$$

and

$$\beta_{\mathrm{I}} := \frac{\log(3C_{\mathrm{I}})}{(1 - 2\varepsilon)\Delta_{\mathrm{gap}}} = \frac{\log(6S)}{(1 - 2\varepsilon)\Delta_{\mathrm{gap}}}.$$

Define $V_{\mathrm{I}}^\star, B_{\mathrm{I}}^\star \in \mathbb{R}^{p \times p}$ by

$$V_{\mathrm{I}}^\star x := (0_d, 0_d, 0, 0, 0_d, x_2),$$

and

$$B_{\mathrm{I}}^\star x := (0_d, 0_d, -2\varepsilon\beta_{\mathrm{I}} x_4, 0, \beta_{\mathrm{I}} x_2, 0_d).$$

The operator norms satisfy

$$\|V_{\mathrm{I}}^\star\|_{\mathrm{op}} = 1, \qquad \|B_{\mathrm{I}}^\star\|_{\mathrm{op}} \le \beta_{\mathrm{I}},$$

because $0 < 2\varepsilon < 1$.

Let

$$Z^{(k)}(\mathcal{T}) := \mathcal{F}_{\psi_{\mathrm{prep}}}(E_{\mathrm{I},k}(\mathcal{T}))$$

be the preprocessed prefix. At teacher-forced step k, the last column is the preprocessed selected-state token

$$z_{\mathrm{cur}}^{(k)} = (0_d, a_{u_k}, 0, 1, 0_d, 0_d).$$

For an edge-memory source column corresponding to $e = (u, v)$, Lemma B.1 gives

$$z_e = (a_u, a_v, r_{\mathcal{T}}(v) + \tau_{\mathrm{I}}, 1, (r_{\mathcal{T}}(v) + \tau_{\mathrm{I}})a_u, 0_d).$$

Therefore its score against the current query is

$$q_k(e) := z_e^\top B_{\mathrm{I}}^\star z_{\mathrm{cur}}^{(k)} = \beta_{\mathrm{I}}(r_{\mathcal{T}}(v) + \tau_{\mathrm{I}})\left(\langle a_u, a_{u_k}\rangle - 2\varepsilon\right).$$

For a selected-state source column z_{st}, the third and fifth blocks are zero, so

$$q_k(z_{\mathrm{st}}) = 0.$$

Lemma B.2 (Softmax tail for the greedy-selection head)**.** *Let*

$$e_k^\star = (u_k, u_{k+1}), \qquad u_{k+1} = \arg\max_{v \in \mathrm{ch}(u_k)} r_{\mathcal{T}}(v).$$

At every teacher-forced greedy step k,

$$\frac{\sum_{i \neq e_k^\star} \exp(q_k(i))}{\exp(q_k(e_k^\star))} \leq \frac{1}{3}.$$

Consequently, if $\omega_k^\star$ denotes the softmax weight on the oracle edge $e_k^\star$, then

$$1 - \omega_k^\star \leq \frac{1}{4}.$$

Proof. For the oracle edge $e_k^\star = (u_k, u_{k+1})$,

$$q_k(e_k^\star) = \beta_{\mathrm{I}}(r_{\mathcal{T}}(u_{k+1}) + \tau_{\mathrm{I}})(1 - 2\varepsilon).$$

Let $\kappa_\varepsilon = 1 - 2\varepsilon$.

First consider another outgoing edge (u_k, v) with $v \neq u_{k+1}$. By the reward-gap assumption,

$$r_{\mathcal{T}}(u_{k+1}) - r_{\mathcal{T}}(v) \geq \Delta_{\mathrm{gap}}.$$

Since the parent embedding matches in both scores,

$$q_k(e_k^\star) - q_k(u_k, v) = \beta_{\mathrm{I}}\kappa_\varepsilon\left(r_{\mathcal{T}}(u_{k+1}) - r_{\mathcal{T}}(v)\right) \geq \beta_{\mathrm{I}}\kappa_\varepsilon\Delta_{\mathrm{gap}}.$$

Next consider an edge (u, v) with $u \neq u_k$. By near-orthogonality,

$$\langle a_u, a_{u_k}\rangle \leq \varepsilon.$$

Also

$$r_{\mathcal{T}}(v) + \tau_{\mathrm{I}} \geq \tau_{\mathrm{I}} - R = \Delta_{\mathrm{gap}} > 0.$$

Hence

$$q_k(u, v) \leq -\beta_{\mathrm{I}}\varepsilon(r_{\mathcal{T}}(v) + \tau_{\mathrm{I}}) \leq 0.$$

On the other hand,

$$q_k(e_k^\star) \geq \beta_{\mathrm{I}}\kappa_\varepsilon(\tau_{\mathrm{I}} - R) = \beta_{\mathrm{I}}\kappa_\varepsilon\Delta_{\mathrm{gap}}.$$

Thus

$$q_k(e_k^\star) - q_k(u, v) \geq \beta_{\mathrm{I}}\kappa_\varepsilon\Delta_{\mathrm{gap}}.$$

Finally, every selected-state source column has score 0. Therefore the same lower bound holds:

$$q_k(e_k^\star) - q_k(z_{\mathrm{st}}) \geq \beta_{\mathrm{I}}\kappa_\varepsilon\Delta_{\mathrm{gap}}.$$

It remains only to count source columns. Since $|\mathcal{V}(\mathcal{T})| \leq S$, the ordinary edge set satisfies

$$|\mathcal{E}(\mathcal{T})| \leq S - 1.$$

Problem I follows a simple root-to-leaf path, so the greedy trajectory length satisfies

$$L_{\mathrm{I}}(\mathcal{T}) \le S - 1.$$

At prediction step k, the prefix contains $k+1$ selected-state tokens, and $k+1 \le L_{\mathrm{I}}(\mathcal{T}) \le S-1$. Hence the number of source columns is at most

$$|\mathcal{E}(\mathcal{T})| + (k+1) \le 2S - 2 \le C_{\mathrm{I}}.$$

Therefore

$$\frac{\sum_{i \neq e_k^\star} \exp(q_k(i))}{\exp(q_k(e_k^\star))} \le C_{\mathrm{I}} \exp\left(-\beta_{\mathrm{I}} \kappa_\varepsilon \Delta_{\mathrm{gap}}\right) = C_{\mathrm{I}} \exp\left(-\log(3C_{\mathrm{I}})\right) = \frac{1}{3}.$$

Since

$$1 - \omega_k^\star = \frac{\sum_{i \neq e_k^\star} \exp(q_k(i)) / \exp(q_k(e_k^\star))}{1 + \sum_{i \neq e_k^\star} \exp(q_k(i)) / \exp(q_k(e_k^\star))},$$

we obtain

$$1 - \omega_k^\star \le \frac{1/3}{1 + 1/3} = \frac{1}{4}.$$

□

The final rounding FFN. For $\mu_{\mathrm{I}} = 1/2$, define the coordinatewise sign-rounder

$$q_{\mu_{\mathrm{I}}}(t) := -\frac{1}{\sqrt{d}} + \frac{1}{\mu_{\mathrm{I}}} \sigma\left(t + \frac{\mu_{\mathrm{I}}}{\sqrt{d}}\right) - \frac{1}{\mu_{\mathrm{I}}} \sigma\left(t - \frac{\mu_{\mathrm{I}}}{\sqrt{d}}\right),$$

and

$$Q_{\mu_{\mathrm{I}}}(z) := \left(q_{\mu_{\mathrm{I}}}(z_1), \ldots, q_{\mu_{\mathrm{I}}}(z_d)\right)^\top.$$

The final FFN is the columnwise lift of

$$\psi_{\mathrm{round}}(x) := (0_d, Q_{\mu_{\mathrm{I}}}(x_6), 0, 0, 0_d, 0_d).$$

Lemma B.3 (Exact rounding after greedy selection)**.** *For every $\mathcal{T}$ and every teacher-forced step k,*

$$\psi_{\mathrm{round}}\left(\mathcal{A}_{V_{\mathrm{I}}^\star, B_{\mathrm{I}}^\star}\left(Z^{(k)}(\mathcal{T})\right)_{:,-1}\right) = s_{\mathrm{I}}(u_{k+1}).$$

Proof. The residual attention output at the last column has sixth block

$$\bar{a}_k = \sum_i \omega_{k,i} \, (z_i)_2,$$

where z_i ranges over the preprocessed source columns and $\omega_{k,i}$ are the corresponding softmax weights. For the oracle edge $e_k^\star = (u_k, u_{k+1})$,

$$(z_{e_k^\star})_2 = a_{u_{k+1}}.$$

All source second blocks are sign embeddings, hence have coordinatewise norm at most $1/\sqrt{d}$. By Lemma B.2,

$$1 - \omega_k^\star \le \frac{1}{4}.$$

Therefore, for every coordinate,

$$\left\|\bar{a}_k - a_{u_{k+1}}\right\|_\infty \le \frac{2}{\sqrt{d}}(1 - \omega_k^\star) \le \frac{1}{2\sqrt{d}} = \frac{1 - \mu_{\mathrm{I}}}{\sqrt{d}}.$$

By coordinatewise sign-rounding,

$$Q_{\mu_{\mathrm{I}}}(\bar{a}_k) = a_{u_{k+1}}.$$

The final FFN places this rounded vector in the second block and sets all other blocks to zero. Hence

$$\psi_{\mathrm{round}}\left(\mathcal{A}_{V_{\mathrm{I}}^\star, B_{\mathrm{I}}^\star}\left(Z^{(k)}(\mathcal{T})\right)_{:,-1}\right) = (0_d, a_{u_{k+1}}, 0, 0, 0_d, 0_d) = s_{\mathrm{I}}(u_{k+1}).$$

□

FFN realization and norm bounds. It remains to verify that the two columnwise maps above lie in the norm-bounded FFN class.

Lemma B.4 (ReLU realization of the Problem I FFNs)**.** *There exist universal constants $C_W, C_\Lambda, C_c > 0$ such that the maps ψ_{prep} and ψ_{round} can be realized by ReLU FFNs with depth at least 2, width at most $C_W d$, and bounds*

$$\max_\ell \{\|A_\ell\|_{\mathrm{op}}, \|b_\ell\|_\infty\} \le C_\Lambda \sqrt{M_{\mathrm{I}}},$$

and

$$\|\psi_{\mathrm{prep}}(x)\|_2 \le C_c M_{\mathrm{I}} \|x\|_2, \qquad \|\psi_{\mathrm{round}}(x)\|_2 \le C_c \|x\|_2 \le C_c M_{\mathrm{I}} \|x\|_2.$$

Proof. Both maps are explicit finite ReLU combinations.

The identity copies x_1 and x_2, and the linear registers $x_3 + \tau_{\mathrm{I}} x_4$ and x_4, are implemented by

$$t = \sigma(t) - \sigma(-t).$$

The state detector C_{st} uses the $2d$ ReLU units

$$\sigma(\sqrt{d}\, x_{2,\ell} - x_4), \qquad \sigma(-\sqrt{d}\, x_{2,\ell} - x_4), \qquad \ell = 1, \dots, d.$$

The synthesis register $P(x)$ uses $4d$ ReLU units,

$$\sigma(a_\tau(x) + \lambda_{\mathrm{I}} s_\ell(x)), \quad \sigma(\lambda_{\mathrm{I}} s_\ell(x)), \quad \sigma(a_\tau(x) - \lambda_{\mathrm{I}} s_\ell(x)), \quad \sigma(-\lambda_{\mathrm{I}} s_\ell(x)),$$

for $\ell = 1, \dots, d$. Hence ψ_{prep} has width $O(d)$.

The only coefficients depending on the problem scale are τ_{I} and λ_{I}, both $O(M_{\mathrm{I}})$. Using the positive homogeneity of ReLU, each group of ReLU units can be balanced so that the largest layer norm is $O(\sqrt{M_{\mathrm{I}}})$. The resulting preprocessing map is homogeneous at the origin and has Lipschitz scale $O(M_{\mathrm{I}})$, giving

$$\|\psi_{\mathrm{prep}}(x)\|_2 \le C M_{\mathrm{I}} \|x\|_2.$$

For the final rounder, since $\mu_{\mathrm{I}} = 1/2$, each coordinate is implemented by

$$q_{\mu_{\mathrm{I}}}(t) = -\frac{1}{\sqrt{d}} + 2\sigma\left(t + \frac{1}{2\sqrt{d}}\right) - 2\sigma\left(t - \frac{1}{2\sqrt{d}}\right).$$

Thus $Q_{\mu_{\mathrm{I}}}$ uses $2d$ ReLU units and all weights and biases are bounded by a universal constant. Moreover $q_{\mu_{\mathrm{I}}}(0) = 0$ and $|q_{\mu_{\mathrm{I}}}(t)| \le 2|t|$. Hence

$$\|\psi_{\mathrm{round}}(x)\|_2 = \|Q_{\mu_{\mathrm{I}}}(x_6)\|_2 \le 2\|x_6\|_2 \le 2\|x\|_2.$$

If $L_{\mathrm{ffn}} > 2$, exact identity ReLU layers are inserted using $y = \sigma(y) - \sigma(-y)$, with width $O(d)$. Enlarging the universal constants absorbs these layers. This proves the claim. □

Proof of Theorem 3.1. Define the reference controller

$$F_{\mathrm{I}}^\star := \mathcal{F}_{\psi_{\mathrm{round}}} \circ \mathcal{A}_{V_{\mathrm{I}}^\star, B_{\mathrm{I}}^\star} \circ \mathcal{F}_{\psi_{\mathrm{prep}}}.$$

By Lemma B.4, the FFNs belong to the specified FFN class whenever

$$L_{\mathrm{ffn}} \ge 2, \qquad W_{\mathrm{ffn}} \gtrsim d,$$

$$\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1 + R + \Delta_{\mathrm{gap}}}, \qquad c_{\mathrm{ffn}} \gtrsim 1 + R + \Delta_{\mathrm{gap}}.$$

Also

$$\|V_{\mathrm{I}}^\star\|_{\mathrm{op}} = 1, \qquad \|B_{\mathrm{I}}^\star\|_{\mathrm{op}} \le \beta_{\mathrm{I}} = \frac{\log(6S)}{(1 - 2\varepsilon)\Delta_{\mathrm{gap}}}.$$

Thus the assumptions

$$\Lambda_V \gtrsim 1, \qquad \Lambda_B \gtrsim \frac{\log(2S)}{(1-2\varepsilon)\Delta_{\mathrm{gap}}}$$

ensure that the attention head is contained in the prescribed attention class. Therefore

$$F_{\mathrm{I}}^{\star} \in \mathcal{F}_{\mathrm{I}}(\mathbf{\Lambda}_{\mathrm{I}})$$

under the budget conditions of Theorem 3.1.

Now fix any

$$\mathcal{T} \in \mathrm{supp}(P_{\mathrm{tr,I}})$$

and any teacher-forced step $k = 0, \ldots, L_{\mathrm{I}}(\mathcal{T}) - 1$. By Lemma B.3,

$$F_{\mathrm{I}}^{\star}\left(E_{\mathrm{I},k}(\mathcal{T})\right)_{:,-1} = s_{\mathrm{I}}(u_{k+1}) = o_{\mathrm{I},k+1}(\mathcal{T}).$$

Hence every summand in the teacher-forced squared loss is zero, and therefore

$$\mathcal{R}_{\mathrm{I}}(F_{\mathrm{I}}^{\star}) = 0.$$

Since the target transition is deterministic, the Bayes risk over all measurable predictors is also zero:

$$\mathcal{R}_{\mathrm{I}}^{\star} = 0.$$

Consequently,

$$\mathcal{E}_{\mathrm{app,I}}(\mathbf{\Lambda}_{\mathrm{I}}) = \inf_{F \in \mathcal{F}_{\mathrm{I}}(\mathbf{\Lambda}_{\mathrm{I}})} \mathcal{R}_{\mathrm{I}}(F) - \mathcal{R}_{\mathrm{I}}^{\star} = 0.$$

This completes the proof. □

Specialization to bounded branching and depth. If the explicit trees have branching factor at most $n \geq 2$ and depth parameter at most $H \geq 1$, with H counting tree levels, then, since $S \leq S_{n,H} := 1 + n + \cdots + n^{H-1} \leq 2n^{H-1}$, we have $\log(2S_{n,H}) \leq (H-1)\log n + \log 4 = O(H \log n)$. Therefore Theorem 3.1 applies with $\Lambda_B = O\Big(\frac{H \log n}{(1-2\varepsilon)\Delta_{\mathrm{gap}}}\Big)$.

B.2. Problem II: reward-ordered DFS on an explicit tree

This subsection proves Theorem 3.2. We use the compact tokenization $p_{\mathrm{II}} = 4d + 6$. The construction is

$$F_{\mathrm{II}}^{\star} = \mathcal{F}_{\psi_{\mathrm{out}}} \circ \mathcal{A}_{\mathrm{DFS}} \circ \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{id}}},$$

where ψ_{id} is the identity FFN. The first attention layer detects whether a memory-child node has already appeared in the selected-state trajectory. The overwrite FFN converts this soft visited signal into exact candidate registers and prepares reward-weighted parent registers for unvisited memory edges. The second attention layer performs reward-weighted forward selection, backtracking, and fixed-threshold route detection. The final FFN applies route-gated rounding.

Block notation. Write $p = p_{\mathrm{II}} = 4d + 6$ and decompose

$$x = (x_P, x_C, x_W, x_B, x_r, x_m, x_s, x_\eta, x_\chi, x_{\mathrm{rt}}) \in \mathbb{R}^p,$$

where

$$x_P, x_C, x_W, x_B \in \mathbb{R}^d$$

and

$$x_r, x_m, x_s, x_\eta, x_\chi, x_{\mathrm{rt}} \in \mathbb{R}.$$

The intended meanings are as follows:

x_P = parent block, x_C = memory child before overwrite, and common node block after overwrite,

x_W = state current before overwrite, and forward summary after DFS attention, x_B = reward-weighted parent register before DFS att

$$x_r = \text{reward register}, \quad x_m = \text{memory indicator}, \quad x_s = \text{state indicator},$$
$$x_\eta = \text{soft unvisited signal}, \quad x_\chi = \text{exact unvisited flag}, \quad x_{\mathrm{rt}} = \text{route scalar}.$$

The input memory and selected-state tokens are

$$m_{\mathrm{II}}(u, v) = (a_u, a_v, 0_d, 0_d, r_{\mathcal{T}}(v), 1, 0, 0, 0, 0),$$

and

$$s_{\mathrm{II}}(p, u) = (a_p, 0_d, a_u, 0_d, 0, 0, 1, 0, 0, 0).$$

Internal constants. Let

$$\kappa_\varepsilon := 1 - 2\varepsilon, \qquad M_{\mathrm{II}} := 1 + R + \Delta_{\mathrm{gap}}.$$

We fix the construction-specific constants

$$\tau_{\mathrm{II}} := R + \Delta_{\mathrm{gap}}, \qquad \mu_{\mathrm{II}} := \frac{1}{2}, \qquad \delta_{\mathrm{vis}} := \frac{1}{8}, \qquad \delta_{\mathrm{rt}} := \frac{1}{8}.$$

Let

$$C_{\mathrm{II}} := 4S.$$

Since the explicit tree has at most S ordinary nodes, the augmented memory contains at most S edge tokens, and a full DFS walk contains at most $2S - 1$ selected-state tokens. Hence every teacher-forced prefix contains $O(S)$ source columns, and C_{II} is a uniform source-count upper bound.

Choose inverse-temperature scales satisfying

$$\alpha_{\mathrm{v}} \gtrsim \frac{\log(8C_{\mathrm{II}})}{\kappa_\varepsilon}, \qquad \alpha_{\mathrm{b}} \gtrsim \frac{\log(8C_{\mathrm{II}})}{\kappa_\varepsilon},$$
$$\alpha_{\mathrm{r}} \gtrsim \frac{\log(8C_{\mathrm{II}})}{1 - \varepsilon}, \qquad \beta_{\mathrm{f}} \gtrsim \frac{\log(8C_{\mathrm{II}})}{\kappa_\varepsilon \Delta_{\mathrm{gap}}}.$$

The constants are chosen large enough for the tail bounds below. Since $\varepsilon < 1/2$, all these scales are bounded by

$$O\left(\frac{\log(2S)}{1 - 2\varepsilon} \max\{1, \Delta_{\mathrm{gap}}^{-1}\}\right).$$

Visited-state detection. The visited detector uses one head. In the $J = 3$ architecture, the two unused heads in the first attention layer are set to zero. Define the visited value map by

$$V_{\mathrm{vis}} y = (0_d, 0_d, 0_d, 0_d, 0, 0, 0, y_m, 0, 0),$$

so that the head writes the total attention mass on memory tokens into the x_η register.

For source y and query x, define the visited score by

$$q_{\mathrm{vis}}(y, x) := \alpha_{\mathrm{v}} \left(\langle y_W, x_C \rangle - \theta_{\mathrm{v}} y_s\right), \qquad \theta_{\mathrm{v}} := \frac{1 + 2\varepsilon}{2}.$$

Equivalently, this score is realized by a matrix $B^\star_{\mathrm{vis}}$ satisfying

$$y^\top B^\star_{\mathrm{vis}} x = q_{\mathrm{vis}}(y, x).$$

The memory-token queries use this head to detect whether their child block has appeared as a previous selected-state current node.

Lemma B.5 (Visited detector)**.** *Let $m_{\mathrm{II}}(u, v)$ be a memory query at teacher-forced step k. After applying the visited attention layer, its x_η-register satisfies*

$$x_\eta \geq 1 - \delta_{\mathrm{vis}} \qquad \textit{if } v \notin \mathsf{V}_k,$$

and

$$x_\eta \leq \delta_{\mathrm{vis}} \qquad \textit{if } v \in \mathsf{V}_k.$$

Proof. For a memory query $x = m_{\mathrm{II}}(u, v)$, every memory source has score zero because $y_W = 0_d$ and $y_s = 0$. Every selected-state source $s_{\mathrm{II}}(p_t, u_t)$ has score

$$\alpha_{\mathrm{v}} \left(\langle a_{u_t}, a_v \rangle - \theta_{\mathrm{v}} \right).$$

If $v \notin \mathsf{V}_k$, then no selected-state current embedding equals a_v. Thus

$$\langle a_{u_t}, a_v \rangle \le \varepsilon$$

for every state source, and hence every state-source score is at most

$$\alpha_{\mathrm{v}}(\varepsilon - \theta_{\mathrm{v}}) \le -\frac{\alpha_{\mathrm{v}}}{2}.$$

Memory sources have value 1 and state sources have value 0. Since the number of sources is at most C_{II}, the total mass on value-zero sources is at most δ_{vis} for the chosen α_{v}. Therefore $x_\eta \ge 1 - \delta_{\mathrm{vis}}$.

If $v \in \mathsf{V}_k$, there exists a selected-state source with current block a_v. Its score is

$$\alpha_{\mathrm{v}}(1 - \theta_{\mathrm{v}}) = \frac{\alpha_{\mathrm{v}}}{2}(1 - 2\varepsilon) = \frac{\alpha_{\mathrm{v}} \kappa_\varepsilon}{2}.$$

All memory sources still have score zero and value 1, while this matching state source has value 0. Hence the mass on memory sources is at most δ_{vis} by the choice of α_{v}. Therefore $x_\eta \le \delta_{\mathrm{vis}}$. □

Overwrite FFN. The overwrite FFN converts the soft visited signal into an exact unvisited flag and rewrites each token into the representation used by the DFS attention layer. On the margin-separated values produced by Lemma B.5, it implements the following exact map.

For a memory token corresponding to (u, v), define

$$\chi_k(u, v) := \mathbf{1}\{v \notin \mathsf{V}_k\}.$$

Then

$$\psi_{\mathrm{ow}} \left(\mathcal{A}_{\mathrm{vis}}(E_{\mathrm{II},k}(\mathcal{T})) \right)_{(u,v)} = \big(\chi_k(u,v) a_u,\ \chi_k(u,v) a_v,\ 0_d,\ \chi_k(u,v)(r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}) a_u,\ \chi_k(u,v)(r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}),\ 1,\ 0,\ 0,\ \chi_k(u,v),\ 0 \big).$$

For a selected-state token $s_{\mathrm{II}}(p, u)$, it outputs

$$\psi_{\mathrm{ow}}(s_{\mathrm{II}}(p, u)) = (a_p, a_u, 0_d, 0_d, 0, 0, 1, 0, 0, 0).$$

Thus, after overwrite, the x_C-block is a common node block: it is the unvisited child embedding for memory tokens and the current-node embedding for state tokens. The x_B-block stores the reward-weighted parent register for unvisited memory tokens and remains zero on selected-state tokens.

Lemma B.6 (Exact overwrite)**.** *There exists a ReLU FFN ψ_{ow} with width $O(d)$ and depth at least 2 that realizes the map above on every teacher-forced prefix. Moreover, its norm bounds are of order*

$$\Lambda_{\mathrm{ffn}} = O(\sqrt{M_{\mathrm{II}}}), \qquad c_{\mathrm{ffn}} = O(M_{\mathrm{II}}).$$

Proof. The map is a finite combination of ReLU threshold gates and coordinatewise gates. First, since $x_\eta \le \delta_{\mathrm{vis}}$ or $x_\eta \ge 1 - \delta_{\mathrm{vis}}$, a clipped ReLU ramp with threshold $1/2$ gives the exact bit $\chi \in \{0, 1\}$. Multiplication by this bit on the finite sign coordinates is implemented by standard two-ReLU gates. The reward register is first shifted by τ_{II}, and then gated by χ. The reward-weighted parent register $\chi(r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}) a_u$ is implemented coordinatewise using the finite sign values of a_u. Because $|r_{\mathcal{T}}(v)| \le R$ and $\tau_{\mathrm{II}} = R + \Delta_{\mathrm{gap}}$, all scalar quantities involved are $O(M_{\mathrm{II}})$. By positive homogeneity of ReLU, the realization can be balanced so that the layer-norm radius is $O(\sqrt{M_{\mathrm{II}}})$, while the output-domination constant is $O(M_{\mathrm{II}})$. The number of coordinatewise gates is $O(d)$. □

DFS attention layer. Let

$$Z^{(k)}(\mathcal{T}) := \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{id}}} \left(E_{\mathrm{II},k}(\mathcal{T}) \right)$$

be the overwritten prefix. At the last column, corresponding to the current state (p_k, u_k), where $p_k = \mathrm{par}(u_k)$, we have

$$z^{(k)}_{\mathrm{cur}} = (a_{p_k}, a_{u_k}, 0_d, 0_d, 0, 0, 1, 0, 0, 0).$$

The DFS attention layer has three heads.

Forward-selection head. For source y and query x, define

$$q_{\mathrm{f}}(y,x) := \beta_{\mathrm{f}}\left(\langle y_B, x_C\rangle - 2\varepsilon\, y_r x_s - 2 y_s x_s\right).$$

The value map is

$$V_{\mathrm{f}} y = (0_d, 0_d, y_C, 0_d, 0, 0, 0, 0, 0, 0),$$

so this head writes the selected child embedding into the x_W-block.

Backtracking head. For source y and query x, define

$$q_{\mathrm{b}}(y,x) := \alpha_{\mathrm{b}}\left(\langle y_C, x_P\rangle - 2\varepsilon\, y_s x_s - 2 y_m x_s\right).$$

The value map is

$$V_{\mathrm{b}} y = (0_d, 0_d, 0_d, y_P, 0, 0, 0, 0, 0, 0),$$

so this head writes the parent block of the retrieved state into x_B.

Route head. For source y and query x, define the fixed-threshold route score by

$$q_{\mathrm{r}}(y,x) := \alpha_{\mathrm{r}}\left(\langle y_P, x_C\rangle - \theta_{\mathrm{r}} y_\chi x_s - 2 y_s x_s\right), \qquad \theta_{\mathrm{r}} := \frac{1+\varepsilon}{2}.$$

The value map is

$$V_{\mathrm{r}} y = (0_d, 0_d, 0_d, 0_d, 0, 0, 0, 0, 0, y_\chi),$$

so this head writes a scalar route signal into x_{rt}.

The three heads define

$$\mathcal{A}_{\mathrm{DFS}} \in \mathfrak{A}_{p_{\mathrm{II}}}(3, \Lambda_V, \Lambda_B)$$

whenever $\Lambda_V \gtrsim 1$ and Λ_B dominates the scales

$$\beta_{\mathrm{f}}, \alpha_{\mathrm{b}}, \alpha_{\mathrm{r}}.$$

Lemma B.7 (Forward-selection head)**.** *Suppose $A_k^{\mathrm{fwd}} \neq \varnothing$, and let*

$$v_k^\star := \arg\max_{v \in A_k^{\mathrm{fwd}}} r_{\mathcal{T}}(v).$$

Then the x_W-block of

$$\mathcal{A}_{\mathrm{DFS}}(Z^{(k)}(\mathcal{T}))_{:,-1}$$

satisfies

$$\left\|x_W - a_{v_k^\star}\right\|_\infty \le \frac{1}{2\sqrt{d}}.$$

Proof. For an unvisited memory edge (u, v), the overwritten source has

$$y_P = a_u, \qquad y_C = a_v, \qquad y_B = (r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}) a_u, \qquad y_r = r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}, \qquad y_m = 1.$$

The current query has $x_C = a_{u_k}$ and $x_s = 1$. Thus its forward score against the current query is

$$\beta_{\mathrm{f}}(r_{\mathcal{T}}(v) + \tau_{\mathrm{II}})\left(\langle a_u, a_{u_k}\rangle - 2\varepsilon\right).$$

For the oracle edge $(u_k, v_k^\star)$, this is

$$\beta_{\mathrm{f}}(r_{\mathcal{T}}(v_k^\star) + \tau_{\mathrm{II}})(1 - 2\varepsilon).$$

Since

$$r_{\mathcal{T}}(v_k^\star) + \tau_{\mathrm{II}} \ge \tau_{\mathrm{II}} - R = \Delta_{\mathrm{gap}},$$

the oracle score is at least

$$\beta_{\mathrm{f}} \kappa_\varepsilon \Delta_{\mathrm{gap}}.$$

First consider another unvisited outgoing child (u_k, v) with $v \neq v_k^\star$. By reward separation,

$$r_{\mathcal{T}}(v_k^\star) - r_{\mathcal{T}}(v) \geq \Delta_{\mathrm{gap}}.$$

Since the parent embedding matches in both scores,

$$q_{\mathrm{f}}(u_k, v_k^\star) - q_{\mathrm{f}}(u_k, v) = \beta_{\mathrm{f}}\kappa_\varepsilon \left(r_{\mathcal{T}}(v_k^\star) - r_{\mathcal{T}}(v)\right) \geq \beta_{\mathrm{f}}\kappa_\varepsilon \Delta_{\mathrm{gap}}.$$

Next consider an unvisited memory edge (u, v) with $u \neq u_k$. By near-orthogonality,

$$\langle a_u, a_{u_k} \rangle \leq \varepsilon.$$

Also

$$r_{\mathcal{T}}(v) + \tau_{\mathrm{II}} \geq \Delta_{\mathrm{gap}} > 0.$$

Therefore

$$q_{\mathrm{f}}(u, v) \leq -\beta_{\mathrm{f}}\varepsilon(r_{\mathcal{T}}(v) + \tau_{\mathrm{II}}) \leq 0.$$

Thus the oracle edge beats every unvisited off-parent edge by at least

$$\beta_{\mathrm{f}}\kappa_\varepsilon \Delta_{\mathrm{gap}}.$$

Visited memory tokens have $y_B = 0_d$, $y_C = 0_d$, and $y_r = 0$, hence their forward score is 0. State tokens have $y_B = 0_d$, $y_r = 0$, and $y_s = 1$, and hence receive score $-2\beta_{\mathrm{f}}$. Therefore these tokens are also separated from the oracle by at least

$$\beta_{\mathrm{f}}\kappa_\varepsilon \Delta_{\mathrm{gap}}.$$

The number of source columns is at most C_{II}. By the choice of β_{f}, the total non-oracle softmax mass of the forward head is at most $1/4$. Since all value vectors have coordinatewise norm at most $1/\sqrt{d}$, the convex-combination error in the x_W-block is at most

$$\frac{2}{\sqrt{d}} \cdot \frac{1}{4} = \frac{1}{2\sqrt{d}}.$$

□

Lemma B.8 (Backtracking head)**.** *Suppose $A_k^{\mathrm{fwd}} = \varnothing$ and $u_k \neq \mathrm{root}(\mathcal{T})$. Let $p_k = \mathrm{par}(u_k)$. Then the x_B-block of*

$$\mathcal{A}_{\mathrm{DFS}}(Z^{(k)}(\mathcal{T}))_{:,-1}$$

satisfies

$$\left\| x_B - a_{\mathrm{par}(p_k)} \right\|_\infty \leq \frac{1}{2\sqrt{d}}.$$

Proof. The current query has $x_P = a_{p_k}$. Any previous selected-state token whose current node is p_k has

$$y_C = a_{p_k}, \qquad y_P = a_{\mathrm{par}(p_k)}.$$

Such a source receives score

$$\alpha_{\mathrm{b}}(1 - 2\varepsilon).$$

Any selected-state source with current node different from p_k has $\langle y_C, a_{p_k} \rangle \leq \varepsilon$, and hence score at most $-\alpha_{\mathrm{b}}\varepsilon$. Memory sources are penalized by the term $-2\alpha_{\mathrm{b}} y_m x_s$. Therefore the total non-oracle mass is at most $1/4$ for the chosen α_{b}.

There may be multiple selected-state sources with current node p_k, because DFS can revisit a node. However, all such sources have the same parent block $a_{\mathrm{par}(p_k)}$. Thus the oracle value is unique even if the oracle source set is not a singleton. The same convex-combination bound gives the stated coordinatewise error. □

Lemma B.9 (Route head)**.** *The x_{rt}-block of*

$$\mathcal{A}_{\mathrm{DFS}}(Z^{(k)}(\mathcal{T}))_{:,-1}$$

satisfies

$$x_{\mathrm{rt}} \geq 1 - \delta_{\mathrm{rt}} \qquad \text{if } A_k^{\mathrm{fwd}} \neq \varnothing,$$

and

$$x_{\mathrm{rt}} \leq \delta_{\mathrm{rt}} \qquad \text{if } A_k^{\mathrm{fwd}} = \varnothing.$$

Proof. If $A_k^{\mathrm{fwd}} \neq \varnothing$, there is an unvisited memory edge (u_k, v) with $y_\chi = 1$ and $y_P = a_{u_k}$. Its route score is

$$\alpha_{\mathrm{r}}(1 - \theta_{\mathrm{r}}) = \frac{\alpha_{\mathrm{r}}}{2}(1 - \varepsilon).$$

Every value-one unvisited memory edge whose parent is different from u_k has parent match at most ε, and hence score at most

$$\alpha_{\mathrm{r}}(\varepsilon - \theta_{\mathrm{r}}) = -\frac{\alpha_{\mathrm{r}}}{2}(1 - \varepsilon).$$

Visited memory tokens have value 0 and score 0, while state sources are penalized by $-2\alpha_{\mathrm{r}} y_s x_s$. Therefore a value-one matching source dominates all value-zero sources by margin $\alpha_{\mathrm{r}}(1-\varepsilon)/2$, and it dominates all value-one nonmatching sources by margin $\alpha_{\mathrm{r}}(1-\varepsilon)$. By the choice of α_{r}, the route output is at least $1 - \delta_{\mathrm{rt}}$.

If $A_k^{\mathrm{fwd}} = \varnothing$, every value-one unvisited memory edge has parent different from u_k, and hence score at most

$$-\frac{\alpha_{\mathrm{r}}}{2}(1 - \varepsilon).$$

A value-zero source with score zero is available, for example from a visited memory token on the current DFS trajectory. Therefore the mass assigned to value-one sources is at most δ_{rt}, and $x_{\mathrm{rt}} \leq \delta_{\mathrm{rt}}$. □

Final route-gated rounding FFN. The last FFN reads

$$x_P, \quad x_C, \quad x_W, \quad x_B, \quad x_{\mathrm{rt}}$$

from the last column. It outputs a selected-state token. If the route scalar indicates a forward step, it outputs

$$s_{\mathrm{II}}(u_k, v_k^\star).$$

If the route scalar indicates a backtracking step, it outputs

$$s_{\mathrm{II}}(\mathrm{par}(p_k), p_k).$$

The second component of the backtracking token is obtained from the query-side x_P-block, which already stores a_{p_k}.

Lemma B.10 (Exact route-gated rounding)**.** *For every teacher-forced DFS step k,*

$$\mathcal{F}_{\psi_{\mathrm{out}}} \circ \mathcal{A}_{\mathrm{DFS}} \circ \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{id}}} (E_{\mathrm{II},k}(\mathcal{T}))_{:,-1} = o_{\mathrm{II},k+1}(\mathcal{T}).$$

Proof. If $A_k^{\mathrm{fwd}} \neq \varnothing$, then by Lemma B.9,

$$x_{\mathrm{rt}} \geq 1 - \delta_{\mathrm{rt}}.$$

The final FFN therefore takes the forward branch. The next parent embedding is the current embedding $x_C = a_{u_k}$, which is exact. By Lemma B.7,

$$\|x_W - a_{v_k^\star}\|_\infty \leq \frac{1}{2\sqrt{d}}.$$

Since $\mu_{\mathrm{II}} = 1/2$, coordinatewise sign rounding maps x_W exactly to $a_{v_k^\star}$. Thus the output is

$$s_{\mathrm{II}}(u_k, v_k^\star).$$

If $A_k^{\mathrm{fwd}} = \varnothing$, then the step is a backtracking step unless the trajectory has terminated at the root. For all predicted backtracking steps, Lemma B.9 gives

$$x_{\mathrm{rt}} \le \delta_{\mathrm{rt}},$$

so the final FFN takes the backtracking branch. Let $p_k = \mathrm{par}(u_k)$. By Lemma B.8,

$$\left\| x_B - a_{\mathrm{par}(p_k)} \right\|_\infty \le \frac{1}{2\sqrt{d}}.$$

Coordinatewise sign rounding therefore gives exactly $a_{\mathrm{par}(p_k)}$. The current node after backtracking is p_k, and its embedding a_{p_k} is already stored exactly in the query-side x_P-block. Hence the output is

$$s_{\mathrm{II}}(\mathrm{par}(p_k), p_k).$$

This is precisely the teacher target in Problem II. □

Lemma B.11 (FFN realization and norm bounds)**.** *The maps*

$$\psi_{\mathrm{id}}, \qquad \psi_{\mathrm{ow}}, \qquad \psi_{\mathrm{out}}$$

can be realized by ReLU FFNs with depth at least 2, width $O(d)$, and bounds

$$\Lambda_{\mathrm{ffn}} = O\Big(\sqrt{1 + R + \Delta_{\mathrm{gap}}}\Big), \qquad c_{\mathrm{ffn}} = O(1 + R + \Delta_{\mathrm{gap}}).$$

Proof. The identity map is realized by

$$t = \sigma(t) - \sigma(-t)$$

coordinatewise, using $O(d)$ width.

The overwrite map was handled in Lemma B.6. It uses margin-separated thresholding and coordinatewise gates. The only problem-dependent scalar magnitude is $O(1 + R + \Delta_{\mathrm{gap}})$, and the balanced ReLU realization gives layer radius $O(\sqrt{1 + R + \Delta_{\mathrm{gap}}})$ and output-domination constant $O(1 + R + \Delta_{\mathrm{gap}})$.

The final map uses route-gated coordinatewise sign rounding. Since $\delta_{\mathrm{rt}} = 1/8$ and $\mu_{\mathrm{II}} = 1/2$, all route and rounding margins are fixed numerical constants. Thus the route-gated rounding part has constant-size slopes. It is applied coordinatewise to $O(d)$ coordinates, so the width is $O(d)$. Combining this with the output format registers gives the stated FFN bounds. □

Proof of Theorem 3.2. The reference controller is

$$F_{\mathrm{II}}^\star = \mathcal{F}_{\psi_{\mathrm{out}}} \circ \mathcal{A}_{\mathrm{DFS}} \circ \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{id}}}.$$

The first attention layer uses the visited head and two zero heads, and hence belongs to

$$\mathfrak{A}_{p_{\mathrm{II}}}(3, \Lambda_V, \Lambda_B)$$

whenever $\Lambda_V \gtrsim 1$ and

$$\Lambda_B \gtrsim \frac{\log(2S)}{1 - 2\varepsilon}.$$

The DFS attention layer uses the forward, backtracking, and fixed-threshold route heads defined above. Their score matrices are bounded whenever

$$\Lambda_B \gtrsim \frac{\log(2S)}{1 - 2\varepsilon} \max\{1, \Delta_{\mathrm{gap}}^{-1}\}.$$

By Lemma B.11, the three FFNs belong to the prescribed FFN class whenever

$$L_{\mathrm{ffn}} \ge 2, \qquad W_{\mathrm{ffn}} \gtrsim d,$$

$$\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1 + R + \Delta_{\mathrm{gap}}}, \qquad c_{\mathrm{ffn}} \gtrsim 1 + R + \Delta_{\mathrm{gap}}.$$

Thus

$$F_{\mathrm{II}}^\star \in \mathcal{F}_{\mathrm{II}}(\mathbf{\Lambda}_{\mathrm{II}})$$

under the theorem assumptions.

For every $\mathcal{T} \in \mathrm{supp}(P_{\mathrm{tr,II}})$ and every teacher-forced DFS step k, Lemma B.10 gives

$$F^{\star}_{\mathrm{II}}\left(E_{\mathrm{II},k}(\mathcal{T})\right)_{:,-1} = o_{\mathrm{II},k+1}(\mathcal{T}).$$

Hence every summand in the teacher-forced squared loss is zero, and therefore

$$\mathcal{R}_{\mathrm{II}}(F^{\star}_{\mathrm{II}}) = 0.$$

Since the target transition is deterministic, the Bayes risk over all measurable predictors is also zero:

$$\mathcal{R}^{\star}_{\mathrm{II}} = 0.$$

Consequently,

$$\mathcal{E}_{\mathrm{app,II}}(\mathbf{\Lambda}_{\mathrm{II}}) = \inf_{F \in \mathcal{F}_{\mathrm{II}}(\mathbf{\Lambda}_{\mathrm{II}})} \mathcal{R}_{\mathrm{II}}(F) - \mathcal{R}^{\star}_{\mathrm{II}} = 0.$$

This proves the theorem. □

Specialization to bounded branching and depth. If the explicit trees have branching factor at most $n \geq 2$ and depth parameter at most $H \geq 1$, with H counting tree levels, then

$$S \leq S_{n,H} := 1 + n + \cdots + n^{H-1} \leq 2n^{H-1}.$$

Hence

$$\log(2S_{n,H}) = O(H \log n).$$

Therefore Theorem 3.2 applies with

$$\Lambda_B = O\left(\frac{H \log n}{1 - 2\varepsilon} \max\{1, \Delta_{\mathrm{gap}}^{-1}\}\right).$$

In the same setting, the sign-embedding dimension can be chosen as

$$d = O(\varepsilon^{-2} H \log n).$$

B.3. Problem III: DFS control over an implicit generated tree

This subsection proves Theorem 3.3. The fixed front-end is a non-trainable columnwise local proposal map. It reads the source thought, the source-node block, and the one-hot phase, and writes a local proposal

$$(G_z, G_v, G_r) = g_{\mathrm{pre}}(z_{\mathrm{src}}, U, \phi).$$

All search-control operations are implemented by the trainable Transformer controller.

Block notation. Write $p = p_{\mathrm{III}} = 12d + C + 9$ and decompose a token as

$$x = (z, z_{\mathrm{src}}, P, U, V, r, \phi, b, \iota; G_z, G_v, G_r, \chi_{\mathrm{eos}}, \eta_{\mathrm{vis}}, \chi_{\mathrm{unv}}, F_z, F_v, T_{\mathrm{f}}, B_z, B_p, \rho).$$

Here

$$z, z_{\mathrm{src}}, P, U, V, G_z, G_v, F_z, F_v, T_{\mathrm{f}}, B_z, B_p \in \mathbb{R}^d,$$
$$r, b, \iota, G_r, \chi_{\mathrm{eos}}, \eta_{\mathrm{vis}}, \chi_{\mathrm{unv}}, \rho \in \mathbb{R}, \qquad \phi \in \mathbb{R}^{C+1}.$$

The visible blocks P, U, V store, respectively, the parent of the source node, the source/current node, and the candidate child node. The scalar block ι is equal to 1 on legal tokens and is used only to implement the centered Problem I-style forward score. The existing block U is the source-node input to g_{pre}; no additional source-node block is used.

Let $e_1, \ldots, e_{C+1}$ be the standard basis of $\mathbb{R}^{C+1}$. On legal tokens, define the exact phase indicators

$$\mathrm{sel}(\phi) := e_1^{\top}\phi, \qquad \mathrm{cand}(\phi) := \sum_{j=2}^{C+1} e_j^{\top}\phi, \qquad \mathrm{gen}(\phi) := \sum_{j=1}^{C} e_j^{\top}\phi.$$

Thus selected-state tokens have $\phi = e_1$, while the j-th candidate token has $\phi = e_{j+1}$.

The selected-state and candidate tokens are

$$s_{\mathrm{sel}}(z, \pi, u, b) = (z, z, a_\pi, a_u, 0_d, 0, e_1, b, 1; 0_{7d+5}),$$

and

$$s_{\mathrm{cand}}(z', z_{\mathrm{src}}, \pi, u, v, r, j) = (z', z_{\mathrm{src}}, a_\pi, a_u, a_v, r, e_{j+1}, 0, 1; 0_{7d+5}), \qquad j = 1, \dots, C.$$

In particular, selected-state tokens always satisfy $z_{\mathrm{src}} = z$.

Construction constants. Let

$$\kappa_\varepsilon := 1 - 2\varepsilon, \qquad M_{\mathrm{III}} := 1 + R + \Delta_{\mathrm{gap}} + Z, \qquad N_{\mathrm{III}} := (C+2)S.$$

The number N_{III} upper bounds the number of source columns in any teacher-forced Problem III prefix. Fix $\tau_{\mathrm{III}} := R + \Delta_{\mathrm{gap}}$, so

$$r + \tau_{\mathrm{III}} \in [\Delta_{\mathrm{gap}}, 2R + \Delta_{\mathrm{gap}}]$$

for every valid generated reward. Define

$$\gamma_{\mathrm{III}} := \min\{1 - \varepsilon, \kappa_\varepsilon \Delta_{\mathrm{gap}}\}.$$

The inverse temperatures in the construction are chosen as universal constant multiples of the available score radius Λ_B, small enough that all score matrices have operator norm at most Λ_B. Thus the visited, backtracking, and route margins are of order $(1-\varepsilon)\Lambda_B$, while the Problem I-style forward margin is of order $\kappa_\varepsilon \Delta_{\mathrm{gap}} \Lambda_B$. The construction therefore has softmax tails bounded by

$$N_{\mathrm{III}} \exp[-\Omega(\gamma_{\mathrm{III}} \Lambda_B)].$$

In particular, the rounding conditions below hold whenever

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1} \log((C+2)S).$$

Reference controller. The constructed controller has the form

$$F^\star_{\mathrm{III}} = \mathcal{F}_{\psi_{\mathrm{out}}} \circ \mathcal{A}_{\mathrm{DFS}} \circ \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{prep}}} \circ \Phi^\phi_{\mathrm{pre}},$$

where Φ^ϕ_{pre} is the fixed columnwise lift of g_{pre}. This fixed map is not trainable and is not counted in the budget. The first attention layer uses one active visited-detection head and two zero heads. The second attention layer uses three logical heads: forward selection, backtracking retrieval, and route detection.

Fixed proposal layer. The fixed layer Φ^ϕ_{pre} acts columnwise. On a token with visible blocks $(z, z_{\mathrm{src}}, P, U, V, r, \phi, b, \iota)$, it writes

$$(G_z, G_v, G_r) = g_{\mathrm{pre}}(z_{\mathrm{src}}, U, \phi)$$

into the proposal registers and leaves the visible blocks unchanged. It does not decide whether the proposal is used.

Preprocessing FFN. The preprocessing FFN copies all visible and proposal registers and prepares a selected-current register for visited-state detection. On legal tokens it implements

$$F_z \leftarrow \mathrm{sel}(\phi)\, U,$$

and sets the remaining search-control scratch registers to zero. Thus selected tokens write their current node into F_z, while candidate tokens write 0_d.

Lemma B.12 (Preprocessing realization)**.** *The preprocessing map above is realized by a ReLU FFN of depth at least* 2 *and width* $O(d + C)$*. Its layer-norm and output-domination constants are bounded by* $O(\sqrt{M_{\mathrm{III}}})$ *and* $O(M_{\mathrm{III}})$*, respectively.*

Proof. The map consists of coordinatewise identity copies, one-hot phase readouts, and coordinatewise products by the exact bit $\mathrm{sel}(\phi) \in \{0, 1\}$ on legal tokens. These are standard finite ReLU gates. The number of coordinate gates is $O(d)$, and the phase readouts require $O(C)$ width. All copied or gated coordinates have magnitude $O(M_{\mathrm{III}})$, and the usual positive-homogeneity balancing gives the stated norm bounds. □

Visited-state detection. Let

$$H_1 := \mathcal{F}_{\psi_{\mathrm{prep}}} \circ \Phi^{\phi}_{\mathrm{pre}}(E_{\mathrm{III},k}).$$

For a source column y and query column x, the visited head uses the score

$$q_{\mathrm{vis}}(y, x) = \alpha_{\mathrm{vis}} \left(\langle y_{F_z}, x_V \rangle - \theta_{\mathrm{vis}} \mathrm{sel}(y_\phi) \right), \qquad \theta_{\mathrm{vis}} := \frac{1+\varepsilon}{2},$$

and value $\mathrm{sel}(y_\phi)$. The value is written into the η_{vis}-register.

Lemma B.13 (Visited detector for generated candidates)**.** *Let x_i be a candidate token in a teacher-forced prefix, with $V_i = a_v$. Let V_k be the set of ordinary nodes that have appeared as current nodes in selected-state tokens in the prefix. After the visited attention layer,*

$$\eta_{\mathrm{vis},i} \geq 1 - \delta_{\mathrm{vis}} \quad \textit{if } v \in \mathsf{V}_k, \qquad \eta_{\mathrm{vis},i} \leq \delta_{\mathrm{vis}} \quad \textit{if } v \notin \mathsf{V}_k,$$

where

$$\delta_{\mathrm{vis}} \leq N_{\mathrm{III}} \exp[-\Omega((1-\varepsilon)\alpha_{\mathrm{vis}})].$$

Proof. For selected-state sources, $y_{F_z} = y_U$ and $\mathrm{sel}(y_\phi) = 1$. For non-selected sources, $y_{F_z} = 0_d$ and $\mathrm{sel}(y_\phi) = 0$. If $v \in \mathsf{V}_k$, there exists a selected-state source with $y_U = a_v$, whose score is at least

$$\alpha_{\mathrm{vis}}(1 - \theta_{\mathrm{vis}}) = \frac{\alpha_{\mathrm{vis}}}{2}(1-\varepsilon).$$

Every non-selected source has score 0, and every nonmatching selected source has score at most

$$\alpha_{\mathrm{vis}}(\varepsilon - \theta_{\mathrm{vis}}) = -\frac{\alpha_{\mathrm{vis}}}{2}(1-\varepsilon).$$

Hence the matching selected sources dominate the softmax, and the total mass on value-zero sources is at most the displayed tail.

If $v \notin \mathsf{V}_k$, no selected-state source matches a_v. Then all selected-state sources have score at most

$$-\frac{\alpha_{\mathrm{vis}}}{2}(1-\varepsilon),$$

while non-selected sources have score 0 and value 0. Therefore the total mass on value-one sources is at most the same tail bound. □

Overwrite FFN. The overwrite FFN converts the soft visited signal into exact binary registers and prepares masked registers for the DFS attention layer. On teacher-forced prefixes it implements

$$\chi_{\mathrm{eos}} = \mathbf{1}\{G_v = a_{\mathrm{eos}}\},$$

and

$$\chi_{\mathrm{unv}} = \mathbf{1}\{\mathrm{cand}(\phi) = 1\}\mathbf{1}\{V \neq a_{\mathrm{eos}}\}\mathbf{1}\{\eta_{\mathrm{vis}} \leq 1/2\}.$$

It also writes the following temporary scoring registers:

$$F_v \leftarrow \chi_{\mathrm{unv}}\, U, \qquad T_{\mathrm{f}} \leftarrow \chi_{\mathrm{unv}}(r + \tau_{\mathrm{III}})U, \qquad B_p \leftarrow \mathrm{sel}(\phi)\, U, \qquad \rho \leftarrow \chi_{\mathrm{unv}}(r + \tau_{\mathrm{III}}).$$

All other temporary registers used by the next attention layer are reset to zero, while the visible blocks, the constant block ι, and the proposal registers $(G_z, G_v, G_r, \chi_{\mathrm{eos}})$ are copied.

Lemma B.14 (Overwrite realization)**.** *Assume that the visited detector satisfies $\delta_{\mathrm{vis}} \leq 1/4$. The overwrite map above is realized exactly on all teacher-forced Problem III prefixes by a ReLU FFN of depth at least 2 and width $O(d + C)$. Its layer-norm and output-domination constants are bounded by $O(\sqrt{M_{\mathrm{III}}})$ and $O(M_{\mathrm{III}})$, respectively.*

Proof. The EOS test is a margin threshold on $\langle G_v, a_{\mathrm{eos}} \rangle$. Since node embeddings are near-orthogonal and $\varepsilon < 1/2$, the threshold $(1+\varepsilon)/2$ separates $G_v = a_{\mathrm{eos}}$ from all ordinary generated nodes. The visited signal is separated by Lemma B.13, so a threshold at $1/2$ gives the exact bit. The candidate-phase bit is a linear readout of the one-hot phase. Products of exact bits with sign coordinates and bounded scalar rewards are implemented by standard ReLU gates, as in the Problem I preprocessing construction. The width is $O(d + C)$, and all scalar magnitudes are $O(M_{\mathrm{III}})$. Balancing the ReLU realization gives the claimed norm bounds. □

DFS attention layer. Let

$$H_3 := \mathcal{F}_{\psi_{\mathrm{ow}}} \circ \mathcal{A}_{\mathrm{vis}} \circ \mathcal{F}_{\psi_{\mathrm{prep}}} \circ \Phi^{\phi}_{\mathrm{pre}}(E_{\mathrm{III},k}).$$

Let the last column have source/current node U_{cur}, parent P_{cur}, and $\iota = 1$. The DFS attention layer has three heads.

The forward-selection head uses the Problem I-style score

$$q_{\mathrm{f}}(y, x) = \beta_{\mathrm{f}} \left(\langle y_{T_{\mathrm{f}}}, x_U \rangle - 2\varepsilon \, y_\rho x_\iota \right),$$

and writes the source z-block and V-block into F_z and F_v, respectively. Since

$$y_{T_{\mathrm{f}}} = y_{\chi_{\mathrm{unv}}}(y_r + \tau_{\mathrm{III}}) y_U, \qquad y_\rho = y_{\chi_{\mathrm{unv}}}(y_r + \tau_{\mathrm{III}}),$$

and $x_\iota = 1$ on legal queries, only valid unvisited candidate tokens carry a nonzero reward-weighted source register.

The backtracking head uses the score

$$q_{\mathrm{b}}(y, x) = \alpha_{\mathrm{b}} \left(\langle y_{B_p}, x_P \rangle - \theta_{\mathrm{b}} \mathrm{sel}(y_\phi) \right), \qquad \theta_{\mathrm{b}} := \frac{1+\varepsilon}{2},$$

and writes the source z-block and P-block into B_z and B_p, respectively.

The route head uses the score

$$q_{\mathrm{rt}}(y, x) = \alpha_{\mathrm{rt}} \left(\langle y_{F_v}, x_U \rangle - \theta_{\mathrm{rt}} y_{\chi_{\mathrm{unv}}} \right), \qquad \theta_{\mathrm{rt}} := \frac{1+\varepsilon}{2},$$

and writes the value $y_{\chi_{\mathrm{unv}}}$ into the route scalar ρ.

Lemma B.15 (DFS attention summaries)**.** *Let A_k^{fwd} be the set of valid unvisited candidate tokens in the prefix whose source-node block equals U_{cur}. There is a quantity*

$$\delta_{\mathrm{DFS}} \le N_{\mathrm{III}} \exp[-\Omega(\gamma_{\mathrm{III}} \Lambda_B)]$$

such that the following hold at every teacher-forced step.

If $A_k^{\mathrm{fwd}} \neq \emptyset$ and $i^\star = \arg\max_{i \in A_k^{\mathrm{fwd}}} r_i$, then after the DFS attention layer,

$$F_z = (1 - \delta_{\mathrm{f}}) z_{i^\star} + \delta_{\mathrm{f}} \bar{z}_{\mathrm{f}}, \qquad \|F_v - V_{i^\star}\|_\infty \le \frac{2\delta_{\mathrm{f}}}{\sqrt{d}}, \qquad \rho \ge 1 - \delta_{\mathrm{rt}},$$

where $\delta_{\mathrm{f}}, \delta_{\mathrm{rt}} \le \delta_{\mathrm{DFS}}$ and $\|\bar{z}_{\mathrm{f}}\|_2 \le Z$.

If $A_k^{\mathrm{fwd}} = \emptyset$ and the current node is not the root, let the most recent selected-state token with current node $P_{\mathrm{cur}} = a_\pi$ be $s_{\mathrm{sel}}(z_\pi, \pi', \pi, \cdot)$. Then after the DFS attention layer,

$$B_z = (1 - \delta_{\mathrm{b}}) z_\pi + \delta_{\mathrm{b}} \bar{z}_{\mathrm{b}}, \qquad \|B_p - a_{\pi'}\|_\infty \le \frac{2\delta_{\mathrm{b}}}{\sqrt{d}}, \qquad \rho \le \delta_{\mathrm{rt}},$$

where $\delta_{\mathrm{b}}, \delta_{\mathrm{rt}} \le \delta_{\mathrm{DFS}}$ and $\|\bar{z}_{\mathrm{b}}\|_2 \le Z$.

Proof. Consider the forward head. For a valid unvisited candidate from the current source, the score is

$$\beta_{\mathrm{f}}(r + \tau_{\mathrm{III}})(1 - 2\varepsilon).$$

Among valid unvisited candidates from the current source, the reward gap gives a score margin at least

$$\beta_{\mathrm{f}}(1 - 2\varepsilon)\Delta_{\mathrm{gap}}.$$

For a valid unvisited candidate from a different source, near-orthogonality gives $\langle y_U, x_U \rangle \le \varepsilon$, and hence its score is at most

$$-\beta_{\mathrm{f}}\varepsilon(r + \tau_{\mathrm{III}}) \le 0.$$

Invalid or already visited candidates have $\chi_{\mathrm{unv}} = 0$, hence $T_{\mathrm{f}} = 0$ and $\rho = 0$, so their forward score is 0. Since the oracle valid candidate has shifted reward at least Δ_{gap}, its score is at least

$$\beta_{\mathrm{f}}(1 - 2\varepsilon)\Delta_{\mathrm{gap}}.$$

Therefore the total non-oracle mass of the forward head is at most δ_{DFS}. The displayed bounds on F_z and F_v follow because all thought vectors have norm at most Z and all node embeddings have coordinatewise norm $1/\sqrt{d}$.

For the backtracking head, the source register B_p is equal to U on selected-state sources and zero on non-selected sources. A selected-state source whose current node is P_{cur} receives score at least

$$\alpha_{\mathrm{b}}(1 - \theta_{\mathrm{b}}) = \frac{\alpha_{\mathrm{b}}}{2}(1 - \varepsilon).$$

Nonmatching selected-state sources receive score at most

$$\alpha_{\mathrm{b}}(\varepsilon - \theta_{\mathrm{b}}) = -\frac{\alpha_{\mathrm{b}}}{2}(1 - \varepsilon),$$

and non-selected sources receive score 0. Hence the matching selected-state sources dominate. Multiple matching selected-state tokens, if present, carry the same thought and parent for the generated node, so the retrieved value is unique. This proves the bounds on B_z and B_p.

Finally, the route head uses the centered structural score with value χ_{unv}. If a forward candidate exists, a value-one matching source receives score

$$\alpha_{\mathrm{rt}}(1 - \theta_{\mathrm{rt}}) = \frac{\alpha_{\mathrm{rt}}}{2}(1 - \varepsilon),$$

while value-one nonmatching sources receive score at most

$$\alpha_{\mathrm{rt}}(\varepsilon - \theta_{\mathrm{rt}}) = -\frac{\alpha_{\mathrm{rt}}}{2}(1 - \varepsilon),$$

and value-zero sources have score 0. Hence the route output satisfies $\rho \geq 1 - \delta_{\mathrm{DFS}}$. If no forward candidate exists, all value-one sources have negative score while value-zero sources have score zero, giving $\rho \leq \delta_{\mathrm{DFS}}$. The displayed exponential bound follows from the source-count bound N_{III} and the chosen attention score radius. □

Residual scratch correction. The DFS attention layer is residual. The overwrite FFN placed temporary scoring registers in F_v, B_p, and ρ. Thus the final formatter first removes the known pre-attention temporary contribution from the last column:

$$\widetilde{F}_v := F_v - \chi_{\mathrm{unv}} U, \qquad \widetilde{B}_p := B_p - \mathrm{sel}(\phi) U, \qquad \widetilde{\rho} := \rho - \chi_{\mathrm{unv}}(r + \tau_{\mathrm{III}}).$$

The register T_{f} is used only in the forward score and is ignored by the final formatter. After this correction, $(F_z, \widetilde{F}_v)$ and $(B_z, \widetilde{B}_p)$ are exactly the forward and backtracking summaries from Lemma B.15, and $\widetilde{\rho}$ is the route signal.

Final formatter FFN. The final FFN forms the generation bit

$$m_{\mathrm{gen}} = \mathbf{1}\{b = 0\}\mathbf{1}\{\mathrm{gen}(\phi) = 1\}\mathbf{1}\{\chi_{\mathrm{eos}} = 0\}.$$

If $m_{\mathrm{gen}} = 1$ and $\phi = e_j$, $j \leq C$, it outputs the candidate token

$$s_{\mathrm{cand}}(G_z, z_{\mathrm{src}}, \pi, u, v, G_r, j), \qquad P = a_\pi, \quad U = a_u, \quad G_v = a_v.$$

If $m_{\mathrm{gen}} = 0$ and $\widetilde{\rho} \geq 1/2$, it outputs

$$s_{\mathrm{sel}}(F_z, u, \mathrm{Round}(\widetilde{F}_v), 0), \qquad U = a_u.$$

If $m_{\mathrm{gen}} = 0$ and $\widetilde{\rho} \leq 1/2$, it outputs

$$s_{\mathrm{sel}}(B_z, \mathrm{Round}(\widetilde{B}_p), \pi, 1), \qquad P = a_\pi.$$

Here Round denotes nearest sign-embedding rounding on the node dictionary.

Lemma B.16 (Final formatter and token error)**.** *Assume $\delta_{\mathrm{DFS}} \le 1/4$. Candidate-generation steps are reproduced exactly. On forward-selection and backtracking steps, all node, phase, and binary-flag blocks are rounded exactly. Moreover, on any selection or backtracking step, the squared token error is at most $O(Z^2\delta_{\mathrm{DFS}}^2)$.*

Proof. On a candidate-generation step, the fixed proposal layer has already written the exact local proposal (G_z, G_v, G_r), and the final FFN simply formats it as the candidate token with the shifted one-hot phase. Hence the output is exact.

On a forward-selection step, Lemma B.15 gives

$$\|\widetilde{F}_v - V_{i^\star}\|_\infty \le \frac{2\delta_{\mathrm{DFS}}}{\sqrt{d}}.$$

Since $\delta_{\mathrm{DFS}} \le 1/4$, this lies inside the coordinatewise rounding basin, so

$$\mathrm{Round}(\widetilde{F}_v) = V_{i^\star}.$$

The parent block, phase, and flag are written exactly. The only continuous error is in the thought block. Since

$$F_z = (1-\delta_{\mathrm{f}})z_{i^\star} + \delta_{\mathrm{f}}\bar{z}_{\mathrm{f}}$$

with both vectors of norm at most Z, we have

$$\|F_z - z_{i^\star}\|_2 \le 2Z\delta_{\mathrm{f}}.$$

The selected token contains this thought in both z and z_{src}, so the squared token error is $O(Z^2\delta_{\mathrm{DFS}}^2)$.

The backtracking case is identical. The rounded parent block is exact because

$$\|\widetilde{B}_p - a_{\pi'}\|_\infty \le \frac{2\delta_{\mathrm{DFS}}}{\sqrt{d}},$$

and the only continuous error is the duplicated thought block B_z, whose error is at most $2Z\delta_{\mathrm{b}}$. □

Lemma B.17 (FFN realization and norm bounds)**.** *The maps ψ_{prep}, ψ_{ow}, and ψ_{out} are realizable by ReLU FFNs with depth at least 2, width $O(d+C)$, and bounds*

$$\Lambda_{\mathrm{ffn}} = O\Big(\sqrt{1+R+\Delta_{\mathrm{gap}}+Z}\Big), \qquad c_{\mathrm{ffn}} = O(1+R+\Delta_{\mathrm{gap}}+Z).$$

Proof. The preprocessing and overwrite maps were handled in Lemmas B.12 and B.14. The final formatter consists of one-hot phase gates, binary threshold gates, coordinatewise subtraction of the known residual scratch terms, coordinatewise sign rounding on node blocks, and exact copying of visible and proposal registers. The one-hot phase shift $e_j \mapsto e_{j+1}$ is a linear map on $\mathbb{R}^{C+1}$. The coordinatewise rounder and binary gates have fixed margins, while the copied or subtracted scalar quantities have magnitude $O(M_{\mathrm{III}})$. Hence the width is $O(d+C)$, and balancing gives the displayed norm bounds. □

Proof of Theorem 3.3. By Lemma B.17, the three FFNs are contained in the prescribed FFN class whenever

$$L_{\mathrm{ffn}} \ge 2, \qquad W_{\mathrm{ffn}} \gtrsim d+C,$$
$$\Lambda_{\mathrm{ffn}} \gtrsim \sqrt{1+R+\Delta_{\mathrm{gap}}+Z}, \qquad c_{\mathrm{ffn}} \gtrsim 1+R+\Delta_{\mathrm{gap}}+Z.$$

The visited attention head and the three DFS heads have value-map norms $O(1)$, and the two unused heads in the visited layer are set to zero, so the attention layers lie in

$$\mathfrak{A}_{p_{\mathrm{III}}}(3, \Lambda_V, \Lambda_B)$$

whenever $\Lambda_V \gtrsim 1$ and

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1}\log((C+2)S).$$

Thus

$$F_{\mathrm{III}}^\star \in \mathcal{F}_{\mathrm{III}}(\mathbf{\Lambda}_{\mathrm{III}}).$$

It remains to bound the teacher-forced risk. Candidate-generation steps are exact by Lemma B.16. On each forward-selection or backtracking step, the squared token error is at most $O(Z^2\delta_{\mathrm{DFS}}^2)$. Since the number of predicted steps is at most $L_{\max,\mathrm{III}} \le (C+2)S-1$, and since

$$\delta_{\mathrm{DFS}} \le (C+2)S\exp[-\Omega(\gamma_{\mathrm{III}}\Lambda_B)]\,,$$

with the trivial bound $\delta_{\mathrm{DFS}} \le 1$, we obtain

$$\mathcal{E}_{\mathrm{app,III}}(\mathbf{\Lambda}_{\mathrm{III}}) \lesssim Z^2 L_{\max,\mathrm{III}} \min\{1,\, ((C+2)S)^2\exp[-\Omega(\gamma_{\mathrm{III}}\Lambda_B)]\}\,.$$

This proves the theorem. □

B.4. Source-selection tuning of constructive attention temperatures

The approximation proofs choose sufficiently large inverse temperatures for the constructive attention heads. This subsection records a head-wise optimization statement showing that these temperatures can be tuned by a simple source-selection loss after the FFN preprocessing maps, value maps, and score directions have been fixed as in the constructions. This is especially useful for Problem III, where some heads transport continuous thought payloads. For such heads, direct squared loss on the transported thought vector need not give a monotone one-dimensional temperature objective, because different wrong thought vectors may point in cancelling directions. The source-selection loss avoids this issue by training the head to put attention mass on the correct source columns.

A single constructive head. Consider one active attention head after the relevant preprocessing FFN has been applied. For each teacher-forced prefix a, let the source columns be indexed by $i = 1, \dots, m_a$, with $m_a \le M$. The fixed base score and value of source i are denoted by

$$s_i^{(a)} \in \mathbb{R}, \qquad v_i^{(a)} \in \mathbb{R}^q.$$

The only trainable parameter is a scalar inverse temperature $\lambda \ge 0$, so that the attention weights are

$$\pi_i^{(a)}(\lambda) = \frac{\exp(\lambda s_i^{(a)})}{\sum_{\ell=1}^{m_a}\exp(\lambda s_\ell^{(a)})}.$$

Let $\mathcal{O}_a \subseteq [m_a]$ be the nonempty oracle source set. These are the source columns whose values should be selected by the constructive head on prefix a. Define the oracle mass and non-oracle tail mass by

$$p_a(\lambda) := \sum_{i\in\mathcal{O}_a}\pi_i^{(a)}(\lambda), \qquad \tau_a(\lambda) := 1 - p_a(\lambda) = \sum_{j\notin\mathcal{O}_a}\pi_j^{(a)}(\lambda).$$

The source-selection loss for prefix a is

$$\ell_a^{\mathrm{src}}(\lambda) := -\log p_a(\lambda),$$

and the empirical source-selection loss over n prefixes is

$$\widehat{L}_h^{\mathrm{src}}(\lambda) := \frac{1}{n}\sum_{a=1}^{n}\ell_a^{\mathrm{src}}(\lambda).$$

We assume a uniform oracle-margin condition: there is $\gamma > 0$ such that

$$s_i^{(a)} - s_j^{(a)} \ge \gamma \qquad (i\in\mathcal{O}_a,\ j\notin\mathcal{O}_a,\ a = 1,\dots,n).$$

Let

$$\Gamma := \max_a \max_{i,j\in[m_a]} |s_i^{(a)} - s_j^{(a)}|.$$

The proposition below is independent of the geometry of the transported values $v_i^{(a)}$. The values only enter the final payload-error consequence.

Proposition B.18 (Source-selection tuning for a constructive attention head)**.** *Under the oracle-margin condition above, for every prefix a and every $\lambda \geq 0$,*

$$\tau_a(\lambda) \leq M \exp(-\gamma\lambda).$$

Fix $0 < \zeta \leq 1/2$, and suppose the initialization satisfies the warm-start condition

$$M \exp(-\gamma\lambda_0) \leq \zeta.$$

Then, for all $\lambda \geq \lambda_0$,

$$-\frac{d}{d\lambda}\widehat{L}_h^{\mathrm{src}}(\lambda) \geq \gamma(1-\zeta)\widehat{L}_h^{\mathrm{src}}(\lambda).$$

Moreover, $\widehat{L}_h^{\mathrm{src}}$ is L_{sm}-smooth on $[\lambda_0, \infty)$, with

$$L_{\mathrm{sm}} \leq C\Gamma^2$$

for a universal constant $C > 0$. Hence gradient descent

$$\lambda_{t+1} = \lambda_t - \eta\frac{d}{d\lambda}\widehat{L}_h^{\mathrm{src}}(\lambda_t), \qquad 0 < \eta \leq L_{\mathrm{sm}}^{-1},$$

initialized at λ_0, satisfies $\lambda_{t+1} \geq \lambda_t$ and

$$\widehat{L}_h^{\mathrm{src}}(\lambda_t) \leq \frac{\widehat{L}_h^{\mathrm{src}}(\lambda_0)}{1 + \frac{\eta\gamma^2(1-\zeta)^2}{2} t \widehat{L}_h^{\mathrm{src}}(\lambda_0)}.$$

In particular,

$$\widehat{L}_h^{\mathrm{src}}(\lambda_t) = O(1/t).$$

Furthermore,

$$\frac{1}{n}\sum_{a=1}^{n} \tau_a(\lambda_t) \leq \widehat{L}_h^{\mathrm{src}}(\lambda_t).$$

If, in addition, all oracle sources carry the same value

$$v_i^{(a)} = y_a \qquad (i \in \mathcal{O}_a),$$

and

$$D := \max_a \max_{j \notin \mathcal{O}_a} \|v_j^{(a)} - y_a\|_2 < \infty,$$

then the head output

$$f_a(\lambda) := \sum_{i=1}^{m_a} \pi_i^{(a)}(\lambda) v_i^{(a)}$$

satisfies

$$\|f_a(\lambda) - y_a\|_2 \leq D\,\tau_a(\lambda),$$

and hence

$$\frac{1}{n}\sum_{a=1}^{n} \|f_a(\lambda_t) - y_a\|_2^2 \leq D^2 \widehat{L}_h^{\mathrm{src}}(\lambda_t) = O(D^2/t).$$

If the temperature is constrained to an interval $[\lambda_0, \Lambda]$, the projected gradient update moves monotonically toward Λ until the projection becomes active.

Proof. Fix a prefix a. Let

$$A_a(\lambda) := \sum_{i \in \mathcal{O}_a} \exp(\lambda s_i^{(a)}), \qquad B_a(\lambda) := \sum_{j \notin \mathcal{O}_a} \exp(\lambda s_j^{(a)}).$$

For every $j \notin \mathcal{O}_a$ and every $i \in \mathcal{O}_a$, the margin assumption gives

$$\exp(\lambda s_j^{(a)}) \le \exp(-\gamma\lambda)\exp(\lambda s_i^{(a)}).$$

Averaging this inequality over $i \in \mathcal{O}_a$ gives

$$\exp(\lambda s_j^{(a)}) \le \frac{\exp(-\gamma\lambda)}{|\mathcal{O}_a|} A_a(\lambda).$$

Summing over $j \notin \mathcal{O}_a$ yields

$$B_a(\lambda) \le M\exp(-\gamma\lambda)A_a(\lambda).$$

Therefore

$$\tau_a(\lambda) = \frac{B_a(\lambda)}{A_a(\lambda)+B_a(\lambda)} \le \frac{B_a(\lambda)}{A_a(\lambda)} \le M\exp(-\gamma\lambda).$$

Next we compute the derivative of the source-selection loss. Let

$$\mu_{\mathcal{O},a}(\lambda) := \frac{\sum_{i\in\mathcal{O}_a}\pi_i^{(a)}(\lambda)s_i^{(a)}}{p_a(\lambda)}$$

be the oracle-restricted mean score. If $\tau_a(\lambda) > 0$, define

$$\mu_{\mathcal{N},a}(\lambda) := \frac{\sum_{j\notin\mathcal{O}_a}\pi_j^{(a)}(\lambda)s_j^{(a)}}{\tau_a(\lambda)}$$

as the non-oracle-restricted mean score. If $\tau_a(\lambda) = 0$, the following derivative inequality is trivial. Otherwise, the margin assumption implies

$$\mu_{\mathcal{O},a}(\lambda) - \mu_{\mathcal{N},a}(\lambda) \ge \gamma.$$

The derivative of the oracle mass is

$$p_a'(\lambda) = p_a(\lambda)\tau_a(\lambda)\left(\mu_{\mathcal{O},a}(\lambda) - \mu_{\mathcal{N},a}(\lambda)\right).$$

Thus

$$-\frac{d}{d\lambda}\ell_a^{\mathrm{src}}(\lambda) = \frac{p_a'(\lambda)}{p_a(\lambda)} = \tau_a(\lambda)\left(\mu_{\mathcal{O},a}(\lambda) - \mu_{\mathcal{N},a}(\lambda)\right) \ge \gamma\tau_a(\lambda).$$

For $\lambda \ge \lambda_0$, the tail bound and warm-start condition imply

$$\tau_a(\lambda) \le \zeta.$$

Since

$$-\log(1-\tau) \le \frac{\tau}{1-\tau} \qquad (0 \le \tau < 1),$$

we have

$$\tau_a(\lambda) \ge (1-\tau_a(\lambda))\ell_a^{\mathrm{src}}(\lambda) \ge (1-\zeta)\ell_a^{\mathrm{src}}(\lambda).$$

Therefore

$$-\frac{d}{d\lambda}\ell_a^{\mathrm{src}}(\lambda) \ge \gamma(1-\zeta)\ell_a^{\mathrm{src}}(\lambda).$$

Averaging over $a = 1, \dots, n$ gives

$$-\frac{d}{d\lambda}\widehat{L}_h^{\mathrm{src}}(\lambda) \ge \gamma(1-\zeta)\widehat{L}_h^{\mathrm{src}}(\lambda).$$

It remains to record smoothness. The loss can be written as

$$\ell_a^{\mathrm{src}}(\lambda) = \log\sum_{i=1}^{m_a}\exp(\lambda s_i^{(a)}) - \log\sum_{i\in\mathcal{O}_a}\exp(\lambda s_i^{(a)}).$$

Hence its second derivative is the difference between the score variance under the full softmax distribution and the score variance under the oracle-restricted softmax distribution. Since all scores for prefix a lie in an interval of length at most Γ, both variances are bounded by $O(\Gamma^2)$. Therefore

$$\left|\frac{d^2}{d\lambda^2}\ell_a^{\mathrm{src}}(\lambda)\right| \le C\Gamma^2$$

for a universal constant $C > 0$, and the same smoothness bound holds for $\widehat{L}_h^{\mathrm{src}}$.

Let

$$c_\gamma := \gamma(1-\zeta).$$

For $0 < \eta \le L_{\mathrm{sm}}^{-1}$, the descent lemma gives

$$\widehat{L}_h^{\mathrm{src}}(\lambda_{t+1}) \le \widehat{L}_h^{\mathrm{src}}(\lambda_t) - \frac{\eta}{2}\left(\frac{d}{d\lambda}\widehat{L}_h^{\mathrm{src}}(\lambda_t)\right)^2.$$

Using

$$-\frac{d}{d\lambda}\widehat{L}_h^{\mathrm{src}}(\lambda_t) \ge c_\gamma \widehat{L}_h^{\mathrm{src}}(\lambda_t),$$

we obtain

$$\widehat{L}_h^{\mathrm{src}}(\lambda_{t+1}) \le \widehat{L}_h^{\mathrm{src}}(\lambda_t) - \frac{\eta c_\gamma^2}{2}\left(\widehat{L}_h^{\mathrm{src}}(\lambda_t)\right)^2.$$

Solving this scalar recursion gives

$$\widehat{L}_h^{\mathrm{src}}(\lambda_t) \le \frac{\widehat{L}_h^{\mathrm{src}}(\lambda_0)}{1 + \frac{\eta c_\gamma^2}{2} t \widehat{L}_h^{\mathrm{src}}(\lambda_0)}.$$

Since the derivative is nonpositive throughout $[\lambda_0, \infty)$, the gradient descent iterates are nondecreasing:

$$\lambda_{t+1} \ge \lambda_t.$$

The average tail bound follows from

$$\tau_a(\lambda) \le -\log(1-\tau_a(\lambda)) = \ell_a^{\mathrm{src}}(\lambda).$$

Finally, if all oracle sources carry value y_a, then

$$f_a(\lambda) - y_a = \sum_{j \notin \mathcal{O}_a} \pi_j^{(a)}(\lambda)(v_j^{(a)} - y_a).$$

Thus

$$\|f_a(\lambda) - y_a\|_2 \le D\tau_a(\lambda).$$

Since $\tau_a(\lambda)^2 \le \tau_a(\lambda) \le \ell_a^{\mathrm{src}}(\lambda)$, averaging gives

$$\frac{1}{n}\sum_{a=1}^{n} \|f_a(\lambda_t) - y_a\|_2^2 \le D^2 \widehat{L}_h^{\mathrm{src}}(\lambda_t).$$

The projected-gradient statement follows because the unprojected update is monotone increasing, so projection onto $[\lambda_0, \Lambda]$ only clips the iterate at the upper endpoint. □

Constructive-head instantiations. The preceding proposition applies to the active constructive heads after the corresponding preprocessing FFNs, value maps, and score directions are fixed. The source-selection oracle set $\mathcal{O}_a$ is the set of source columns that carry the value selected by the constructive proof. For selection and retrieval heads, this is the best child or the retrieved previous state. For binary visited or route heads, it is the source class carrying the required binary value and separated by the constructive score margin.

The relevant source-count bounds and margins are summarized below. One may take $M_{\mathrm{I}} = 2S$, $M_{\mathrm{II}} = 4S$, and $M_{\mathrm{III}} = (C+2)S$.

Problem	head	source-selection margin γ_h
I	greedy selection	$(1-2\varepsilon)\Delta_{\mathrm{gap}}$
II	visited detection	$(1-2\varepsilon)/2$
II	forward selection	$(1-2\varepsilon)\Delta_{\mathrm{gap}}$
II	backtracking retrieval	$1-2\varepsilon$
II	route detection	$(1-\varepsilon)/2$
III	visited detection	$(1-\varepsilon)/2$
III	forward selection	$(1-2\varepsilon)\Delta_{\mathrm{gap}}$
III	backtracking retrieval	$(1-\varepsilon)/2$
III	route detection	$(1-\varepsilon)/2$

For Problem III, the forward-selection and backtracking heads also transport continuous thought payloads. In those cases, the transported thought vectors satisfy $\|z\|_2 \le Z$, so the payload diameter in Proposition B.18 is bounded by

$$D \le 2Z.$$

Therefore source-selection tuning gives

$$\frac{1}{n}\sum_{a=1}^{n} \|f_a(\lambda_t) - y_a\|_2^2 = O(Z^2/t)$$

for the transported thought block, after the warm start and within the fixed constructive subfamily. This payload conclusion does not require any nonnegative-alignment condition on wrong thought vectors; it follows only from source-selection tail control.

Interpretation. The proposition is head-wise and conditional on the constructive score directions and preprocessing maps. It does not assert that unconstrained end-to-end gradient descent over all Transformer parameters finds the full controller. Rather, it shows that the large inverse temperatures used in the approximation theorems can be reached by elementary post-training within the one-dimensional temperature subfamilies. The supervision used for this post-training is the source-selection signal available from teacher-forced search trajectories. In Problem III, this gives a clean optimization route for continuous thought transport: the temperature is trained to select the correct source column, and the transported continuous payload is controlled as a consequence of the resulting softmax-tail bound.

C. Excess Risk Details

This appendix proves Theorem 4.1. The proof is generic and applies to all three problems after substituting the corresponding token dimension, trajectory length, norm radii, and approximation error. The fixed local proposal lift G_{pre} in Problem III is treated as a non-trainable preprocessing map and is not counted in the parameter dimension. When the slow- and fast-rate bounds are used on a common event, we apply the two arguments with confidence parameter $\delta/2$ and combine them by a union bound. Replacing $\log(1/\delta)$ by $\log(2/\delta)$ only changes the logarithmic factors hidden in $\widetilde{O}(\cdot)$.

C.1. Envelope and covering estimates

We first record the deterministic bounds used in the empirical-process argument. Fix $j \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ and a budget

$$\mathbf{\Lambda}_j = (K_j, J_j, L_{\mathrm{ffn},j}, W_{\mathrm{ffn},j}, \Lambda_{V,j}, \Lambda_{B,j}, \Lambda_{\mathrm{ffn},j}, c_{\mathrm{ffn},j}).$$

Let $B_{x,j}$ bound the column norm of the input seen by the trainable controller, and let $B_{y,j}$ bound the target-token norm. For Problem III, $B_{x,\mathrm{III}}$ is taken after the fixed lift G_{pre}, which writes $(G_z, G_v, G_r) = g_{\mathrm{pre}}(z_{\mathrm{src}}, U, \phi)$ columnwise.

Lemma C.1 (Controller output envelope)**.** *For every $F \in \mathcal{F}_j(\mathbf{\Lambda}_j)$ and every teacher-forced prefix $E_{j,k}(\mathcal{T})$,*

$$\|F(E_{j,k}(\mathcal{T}))_{:,-1}\|_2 \le B_{\mathrm{out},j} := c_{\mathrm{ffn},j}^{K_j+1}(1 + J_j\Lambda_{V,j})^{K_j} B_{x,j}.$$

Consequently, the per-instance teacher-forced loss satisfies

$$0 \le \ell_{F,j}(\mathcal{T}) := \frac{1}{2}\sum_{k=0}^{L_j(\mathcal{T})-1} \|F(E_{j,k}(\mathcal{T}))_{:,-1} - o_{j,k+1}(\mathcal{T})\|_2^2 \le \mathsf{B}_j(\mathbf{\Lambda}_j),$$

where

$$\mathsf{B}_j(\mathbf{\Lambda}_j) := \frac{L_{\max,j}}{2}\left(B_{\mathrm{out},j} + B_{y,j}\right)^2.$$

Proof. Each residual attention layer maps a matrix with column norm at most B to a matrix with column norm at most $(1+J_j\Lambda_{V,j})B$, because each attention head outputs a convex combination of value-transformed columns and $\|V_h\|_{\mathrm{op}} \le \Lambda_{V,j}$. Each FFN is $c_{\mathrm{ffn},j}$-output dominated. Iterating over K_j attention layers and K_j+1 FFN layers gives the displayed output bound. The loss bound follows from the triangle inequality and $L_j(\mathcal{T}) \le L_{\max,j}$. □

Let

$$\mathcal{L}_j(\mathbf{\Lambda}_j) := \{\ell_{F,j} : F \in \mathcal{F}_j(\mathbf{\Lambda}_j)\}$$

be the induced per-instance loss class.

Lemma C.2 (Parameter count and loss-class entropy)**.** *Let $D_j(\mathbf{\Lambda}_j)$ be the number of trainable scalar parameters in the attention and FFN layers. One may take*

$$D_j(\mathbf{\Lambda}_j) = O\big(K_jJ_jp_j^2 + (K_j+1)L_{\mathrm{ffn},j}(W_{\mathrm{ffn},j}^2 + p_jW_{\mathrm{ffn},j} + p_j)\big).$$

For Problem III, the fixed lift G_{pre} is not counted in D_{III}. Moreover, there is a quantity $\mathcal{A}_j(\mathbf{\Lambda}_j)$, polynomial in $p_j, L_{\max,j}, B_{x,j}, B_{y,j}, K_j, J_j, L_{\mathrm{ffn},j}, W_{\mathrm{ffn},j}$ and the norm radii in $\mathbf{\Lambda}_j$, such that for every $0 < \varepsilon \le \mathsf{B}_j(\mathbf{\Lambda}_j)$,

$$\log \mathcal{N}(\mathcal{L}_j(\mathbf{\Lambda}_j), \varepsilon, \|\cdot\|_\infty) \le CD_j(\mathbf{\Lambda}_j)\log\left(1 + \frac{\mathcal{A}_j(\mathbf{\Lambda}_j)}{\varepsilon}\right),$$

where $C > 0$ is universal.

Proof. Each attention head contains a value matrix and a combined key-query score matrix, both of size $p_j \times p_j$, giving $O(K_jJ_jp_j^2)$ trainable scalars. A depth-$L_{\mathrm{ffn},j}$, width-$W_{\mathrm{ffn},j}$, input-output dimension p_j ReLU FFN has $O(L_{\mathrm{ffn},j}(W_{\mathrm{ffn},j}^2 + p_jW_{\mathrm{ffn},j} + p_j))$ parameters, and there are K_j+1 such FFNs.

It remains to pass from parameter covers to loss covers. Matrix operator-norm balls have volumetric covers of size $(1 + 2\Lambda/\varepsilon)^{O(p_j^2)}$. The same volumetric argument applied to the FFN parameter tuples gives covers of size $(1 + \mathrm{poly}(\mathbf{\Lambda}_j, p_j, L_{\mathrm{ffn},j}, W_{\mathrm{ffn},j})/\varepsilon)^{O(D_{\mathrm{ffn}})}$. On the deterministic ball supplied by Lemma C.1, attention, residual attention, and output-dominated FFNs are Lipschitz in both their inputs and their parameters, with constants polynomial in $p_j, L_{\max,j}, B_{x,j}, B_{y,j}, K_j, J_j, L_{\mathrm{ffn},j}, W_{\mathrm{ffn},j}$ and the norm radii. Propagating these Lipschitz bounds through the finite-depth composition gives a parameter-to-loss Lipschitz constant absorbed into $\mathcal{A}_j(\mathbf{\Lambda}_j)$. Combining the component covers yields the stated entropy bound. □

C.2. Slow-rate excess-risk bound

We prove the first part of Theorem 4.1.

Lemma C.3 (Uniform convergence from entropy)**.** *Let $\mathcal{L}$ be a class of functions bounded in $[0, B]$. Suppose that for some $D, A > 0$,*

$$\log \mathcal{N}(\mathcal{L}, \varepsilon, \|\cdot\|_\infty) \le D\log\left(1 + \frac{A}{\varepsilon}\right) \qquad (0 < \varepsilon \le B).$$

Then, for i.i.d. samples $Z_1, \dots, Z_N$, with probability at least $1-\delta$,

$$\sup_{\ell\in\mathcal{L}}\left|\mathbb{E}\ell(Z) - \frac{1}{N}\sum_{i=1}^N \ell(Z_i)\right| \le CB\sqrt{\frac{D + \log(1/\delta)}{N}}\sqrt{1 + \log\left(1 + \frac{AN}{B}\right)},$$

where $C > 0$ is universal.

Proof. Take an ε_0-net of $\mathcal{L}$ in sup norm with $\varepsilon_0 := B/N$. Hoeffding's inequality and a union bound over the net give, with probability at least $1-\delta$,

$$\sup_{\ell_0\in\mathcal{N}_{\varepsilon_0}}\left|\mathbb{E}\ell_0 - \frac{1}{N}\sum_{i=1}^N \ell_0(Z_i)\right| \le CB\sqrt{\frac{D\log(1 + A/\varepsilon_0) + \log(1/\delta)}{N}}.$$

Every $\ell \in \mathcal{L}$ is within ε_0 in sup norm of some net element, so replacing ℓ_0 by ℓ adds at most $2\varepsilon_0$. Since $\varepsilon_0 = B/N$, this term is absorbed by the displayed bound after increasing the universal constant. □

Proof of the slow-rate part of Theorem 4.1. Apply Lemma C.3 to $\mathcal{L}_j(\mathbf{\Lambda}_j)$ with $B = \mathsf{B}_j(\mathbf{\Lambda}_j)$, and use Lemma C.2. This gives, with probability at least $1-\delta$,

$$\sup_{F \in \mathcal{F}_j(\mathbf{\Lambda}_j)} \left| \mathcal{R}_j(F) - \widehat{\mathcal{R}}_{j,N}(F) \right| \le \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j) \sqrt{\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}} \right).$$

Fix $\eta > 0$, and choose $F_j^\eta \in \mathcal{F}_j(\mathbf{\Lambda}_j)$ such that

$$\mathcal{R}_j(F_j^\eta) - \mathcal{R}_j^\star \le \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \eta.$$

Since $\widehat{F}_{j,N}$ minimizes the empirical risk,

$$\widehat{\mathcal{R}}_{j,N}(\widehat{F}_{j,N}) \le \widehat{\mathcal{R}}_{j,N}(F_j^\eta).$$

Therefore

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j(F_j^\eta) \le 2 \sup_{F \in \mathcal{F}_j(\mathbf{\Lambda}_j)} \left| \mathcal{R}_j(F) - \widehat{\mathcal{R}}_{j,N}(F) \right|.$$

Adding

$$\mathcal{R}_j(F_j^\eta) - \mathcal{R}_j^\star \le \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \eta$$

gives

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j^\star \le \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \eta + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j) \sqrt{\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}} \right).$$

Letting $\eta \downarrow 0$ proves the slow-rate excess-risk bound. □

C.3. Fast-rate excess-risk bound

We now prove the Bernstein-type fast-rate statement. The only additional ingredient is that the per-instance losses are bounded and nonnegative, and the teacher transitions are deterministic, so $\mathcal{R}_j^\star = 0$. Hence, for any loss $\ell \in [0, B]$, $\mathrm{Var}(\ell) \le \mathbb{E}\ell^2 \le B\,\mathbb{E}\ell$.

Lemma C.4 (Relative deviation for bounded nonnegative classes)**.** *Let $\mathcal{L}$ be a class of functions bounded in $[0, B]$ and satisfying*

$$\log \mathcal{N}(\mathcal{L}, \varepsilon, \|\cdot\|_\infty) \le D \log\left(1 + \frac{A}{\varepsilon}\right) \qquad (0 < \varepsilon \le B).$$

Then, with probability at least $1-\delta$, every $\ell \in \mathcal{L}$ satisfies

$$\mathbb{E}\ell \le 2\mathbb{P}_N \ell + CB \frac{D \log(1 + AN/B) + \log(1/\delta)}{N},$$

and also

$$\mathbb{P}_N \ell \le 2\mathbb{E}\ell + CB \frac{D \log(1 + AN/B) + \log(1/\delta)}{N},$$

where $\mathbb{P}_N \ell := N^{-1} \sum_{i=1}^N \ell(Z_i)$, and $C > 0$ is universal.

Proof. The proof is the standard peeling form of Bernstein's inequality for bounded nonnegative losses. For a fixed ℓ, Bernstein's inequality and $\mathrm{Var}(\ell) \le B\mathbb{E}\ell$ imply a relative deviation bound of the form

$$\mathbb{E}\ell \le 2\mathbb{P}_N \ell + CB \frac{t}{N}, \qquad \mathbb{P}_N \ell \le 2\mathbb{E}\ell + CB \frac{t}{N},$$

with probability at least $1 - e^{-t}$. Apply this inequality on dyadic peels of the range of $\mathbb{E}\ell$, and take a union bound over a B/N-net of the loss class. The entropy bound gives $t \asymp D \log(1 + AN/B) + \log(1/\delta)$. Passing from the net to the full class adds $O(B/N)$, which is absorbed into the displayed term. □

Proof of the fast-rate part of Theorem 4.1. Apply Lemma C.4 to $\mathcal{L}_j(\mathbf{\Lambda}_j)$ with $B = \mathsf{B}_j(\mathbf{\Lambda}_j)$, and use Lemma C.2. Fix $\eta > 0$, and choose $F_j^\eta \in \mathcal{F}_j(\mathbf{\Lambda}_j)$ such that

$$\mathcal{R}_j(F_j^\eta) - \mathcal{R}_j^\star \le \mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \eta.$$

On the relative deviation event,

$$\mathcal{R}_j(\widehat{F}_{j,N}) \le 2\widehat{\mathcal{R}}_{j,N}(\widehat{F}_{j,N}) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}\right).$$

By empirical optimality,

$$\widehat{\mathcal{R}}_{j,N}(\widehat{F}_{j,N}) \le \widehat{\mathcal{R}}_{j,N}(F_j^\eta).$$

Applying the second relative inequality to F_j^η gives

$$\widehat{\mathcal{R}}_{j,N}(F_j^\eta) \le 2\mathcal{R}_j(F_j^\eta) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}\right).$$

Combining the last three displays yields

$$\mathcal{R}_j(\widehat{F}_{j,N}) \le 4\mathcal{R}_j(F_j^\eta) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}\right).$$

Since $\mathcal{R}_j^\star = 0$, this implies

$$\mathcal{R}_j(\widehat{F}_{j,N}) - \mathcal{R}_j^\star \le 4\big(\mathcal{E}_{\mathrm{app},j}(\mathbf{\Lambda}_j) + \eta\big) + \widetilde{O}\left(\mathsf{B}_j(\mathbf{\Lambda}_j)\frac{D_j(\mathbf{\Lambda}_j) + \log(1/\delta)}{N}\right).$$

Letting $\eta \downarrow 0$ proves the fast-rate excess-risk bound. □

C.4. Deriving the problem-level excess-risk rates

We briefly justify the simplified rates stated in Section 4.2. Let $d_S := \varepsilon^{-2}\log(S+2)$, $\ell_\delta := \log(1/\delta)$, $M_{\mathrm{ex}} := 1+R+\Delta_{\mathrm{gap}}$, and $M_{\mathrm{imp}} := 1 + R + \Delta_{\mathrm{gap}} + Z$. The near-orthogonal dictionary uses $d \asymp d_S$.

For Problem I, the approximation theorem uses $K = J = 1$, $p_{\mathrm{I}} = O(d_S)$, $W_{\mathrm{ffn,I}} \asymp d_S$, and $L_{\max,\mathrm{I}} \le S - 1$. Taking the feasible budget at the smallest required order gives $D_{\mathrm{I}} = O(d_S^2)$ and $\mathsf{B}_{\mathrm{I}} = \widetilde{O}(SM_{\mathrm{ex}}^6)$. Since $\mathcal{E}_{\mathrm{app,I}} = 0$, Theorem 4.1 gives

$$\mathcal{R}_{\mathrm{I}}(\widehat{F}_{\mathrm{I},N}) - \mathcal{R}_{\mathrm{I}}^\star = \widetilde{O}\left(SM_{\mathrm{ex}}^6 \min\left\{\frac{d_S + \sqrt{\ell_\delta}}{\sqrt{N}}, \frac{d_S^2 + \ell_\delta}{N}\right\}\right).$$

For Problem II, the approximation theorem uses $K = 2$, $J = 3$, $p_{\mathrm{II}} = O(d_S)$, $W_{\mathrm{ffn,II}} \asymp d_S$, and $L_{\max,\mathrm{II}} \le 2S - 1$. Taking the feasible budget at the smallest required order gives $D_{\mathrm{II}} = O(d_S^2)$ and $\mathsf{B}_{\mathrm{II}} = \widetilde{O}(SM_{\mathrm{ex}}^8)$. Since $\mathcal{E}_{\mathrm{app,II}} = 0$,

$$\mathcal{R}_{\mathrm{II}}(\widehat{F}_{\mathrm{II},N}) - \mathcal{R}_{\mathrm{II}}^\star = \widetilde{O}\left(SM_{\mathrm{ex}}^8 \min\left\{\frac{d_S + \sqrt{\ell_\delta}}{\sqrt{N}}, \frac{d_S^2 + \ell_\delta}{N}\right\}\right).$$

For Problem III, $S = \sum_{h=1}^{H} C^{h-1}$, $p_{\mathrm{III}} = 12d + C + 9$, $W_{\mathrm{ffn,III}} \asymp C + d_S$, and $L_{\max,\mathrm{III}} \le (C+2)S - 1$. Hence, up to logarithmic dependence on the attention-score radius,

$$D_{\mathrm{III}} = O((C + d_S)^2)$$

and

$$\mathsf{B}_{\mathrm{III}} = \widetilde{O}((C+2)SM_{\mathrm{imp}}^8).$$

For a fixed Λ_B, Theorem 4.1 gives the slow and fast statistical scales

$$\mathfrak{e}_{\sqrt{N}}(\Lambda_B) = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8\frac{C + d_S + \sqrt{\ell_\delta}}{\sqrt{N}}\right),$$

and

$$\mathfrak{e}_N(\Lambda_B) = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8\frac{(C + d_S)^2 + \ell_\delta}{N}\right),$$

where the hidden factors include only logarithmic dependence on Λ_B. Combining these scales with the approximation bound and then choosing Λ_B as in Appendix C.5 yields the Problem III rate in the main text.

C.5. Balancing Λ_B in Problem III

This subsection justifies the balanced choice of Λ_B used in Section 4.2. Write

$$\gamma_{\mathrm{III}} := \min\{1-\varepsilon, (1-2\varepsilon)\Delta_{\mathrm{gap}}\}.$$

The approximation theorem gives, for universal constants $C_0, c > 0$,

$$\mathcal{E}_{\mathrm{app,III}}(\Lambda_B) \le C_0 Z^2 L_{\max,\mathrm{III}} \min\left\{1,\ ((C+2)S)^2 \exp(-c\gamma_{\mathrm{III}}\Lambda_B)\right\}.$$

Let

$$\bar{\mathfrak{e}}_{\mathrm{III}} := (C+2)SM_{\mathrm{imp}}^8 \min\left\{\frac{C+d_S+\sqrt{\ell_\delta}}{\sqrt{N}}, \frac{(C+d_S)^2+\ell_\delta}{N}\right\}.$$

It is enough to make the exponential branch of the approximation bound at most a constant multiple of $\bar{\mathfrak{e}}_{\mathrm{III}}$. Since $L_{\max,\mathrm{III}} \le (C+2)S-1$, this is ensured by

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1} \log\left(1 + \frac{Z^2((C+2)S)^3}{\bar{\mathfrak{e}}_{\mathrm{III}}}\right).$$

In addition, the approximation construction itself requires

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1} \log((C+2)S).$$

Combining the two requirements yields the sufficient balanced choice

$$\Lambda_B \gtrsim \gamma_{\mathrm{III}}^{-1} \max\left\{\log((C+2)S),\ \log\left(1 + \frac{Z^2((C+2)S)^3}{\bar{\mathfrak{e}}_{\mathrm{III}}}\right)\right\}.$$

With this choice,

$$\mathcal{E}_{\mathrm{app,III}}(\Lambda_B) \lesssim \bar{\mathfrak{e}}_{\mathrm{III}}.$$

The dependence of the entropy bound on the balanced Λ_B enters only logarithmically through $\mathcal{A}_{\mathrm{III}}(\mathbf{\Lambda}_{\mathrm{III}})$, and is therefore absorbed into the $\widetilde{O}(\cdot)$ factor. Hence, with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{III}}(\widehat{F}_{\mathrm{III},N}) - \mathcal{R}^\star_{\mathrm{III}} = \widetilde{O}\left((C+2)SM_{\mathrm{imp}}^8 \min\left\{\frac{C+d_S+\sqrt{\ell_\delta}}{\sqrt{N}}, \frac{(C+d_S)^2+\ell_\delta}{N}\right\}\right).$$

D. Rounded Autoregressive Execution: Proofs

This appendix proves the rounded-rollout statements in Section 5. Throughout this appendix, $j \in \{\mathrm{I}, \mathrm{II}\}$.

D.1. Nearest-neighbor rounding margins

Let

$$\mathcal{A}(\mathcal{T}) := \{a_q : q \in \{0, \mathrm{eos}\} \cup \mathcal{V}(\mathcal{T})\}.$$

By the sign-embedding condition, for any two distinct symbols $q \ne q'$,

$$\|a_q - a_{q'}\|_2^2 = 2 - 2\langle a_q, a_{q'}\rangle \ge 2(1-\varepsilon).$$

Lemma D.1 (Nearest-node rounding basin)**.** *Let $q \in \{0, \mathrm{eos}\} \cup \mathcal{V}(\mathcal{T})$. If*

$$\|z - a_q\|_2^2 < \frac{1-\varepsilon}{2},$$

then

$$\Pi_{\mathcal{T}}(z) = q.$$

Equivalently, the rounded embedding block satisfies

$$a_{\Pi_{\mathcal{T}}(z)} = a_q.$$

Proof. For any $q' \neq q$,

$$\|z - a_{q'}\|_2 \geq \|a_q - a_{q'}\|_2 - \|z - a_q\|_2 > \sqrt{2(1-\varepsilon)} - \sqrt{\frac{1-\varepsilon}{2}} = \sqrt{\frac{1-\varepsilon}{2}}.$$

Thus $\|z - a_{q'}\|_2 > \|z - a_q\|_2$, so q is the unique nearest-neighbor symbol. Hence $\Pi_{\mathcal{T}}(z) = q$, and the embedding written into the rounded token is $a_{\Pi_{\mathcal{T}}(z)} = a_q$. □

Lemma D.2 (Legal-token rounding basin)**.** *For $j \in \{\mathrm{I}, \mathrm{II}\}$, if*

$$\|\mathsf{D}_{j,k,\mathcal{T}}(y) - \mathsf{D}_{j,k,\mathcal{T}}(o_{j,k+1}(\mathcal{T}))\|_2^2 < \delta_{\mathrm{rd},j}^2,$$

then

$$\mathsf{Q}_{j,\mathcal{T},k+1}(y) = o_{j,k+1}(\mathcal{T}).$$

Proof. The squared error of every block selected by $\mathsf{D}_{j,k,\mathcal{T}}$ is bounded by the displayed masked squared error. Since

$$\delta_{\mathrm{rd},j}^2 = \min\{(1-\varepsilon)/2, 1/4\},$$

every active node block lies in the nearest-node rounding basin of Lemma D.1. Thus, for each active node block z, the symbol $\Pi_{\mathcal{T}}(z)$ is the correct node symbol and the embedding written into the rounded token, $a_{\Pi_{\mathcal{T}}(z)}$, is the correct embedding block. The same threshold also gives correct nearest rounding for binary blocks, because the two legal binary values are separated by distance 1. The projector $\mathsf{Q}_{j,\mathcal{T},k+1}$ then formats all inactive blocks and scratch registers to their canonical legal values. For Problems I and II, all search tokens in the rollout theorem are selected-state tokens, so these rounded control blocks and canonical inactive blocks determine the full teacher target token $o_{j,k+1}(\mathcal{T})$. □

D.2. Proof of teacher-forced to autoregressive conversion

Proof of Theorem 5.1. Fix $j \in \{\mathrm{I}, \mathrm{II}\}$ and F. Define the teacher-forced local control error

$$e_{j,k}^{\mathrm{ctrl}}(F, \mathcal{T}) := \mathsf{D}_{j,k,\mathcal{T}}(F(E_{j,k}(\mathcal{T}))_{:,-1}) - \mathsf{D}_{j,k,\mathcal{T}}(o_{j,k+1}(\mathcal{T}))\,.$$

Let

$$\mathcal{G}_j(F, \mathcal{T}) := \left\{\|e_{j,k}^{\mathrm{ctrl}}(F, \mathcal{T})\|_2^2 < \delta_{\mathrm{rd},j}^2 \quad \text{for all } k = 0, \ldots, L_j(\mathcal{T}) - 1\right\}.$$

On $\mathcal{G}_j(F, \mathcal{T})$, the rounded autoregressive rollout matches the teacher trajectory by induction. At $k = 0$, the initial prefix contains the true token $o_{j,0}(\mathcal{T})$, so the autoregressive prefix equals the teacher-forced prefix. If the prefixes agree up to step k, then the model is evaluated on exactly $E_{j,k}(\mathcal{T})$. The event $\mathcal{G}_j(F, \mathcal{T})$ and Lemma D.2 imply that the rounded output is $o_{j,k+1}(\mathcal{T})$. Hence the prefixes agree at step $k + 1$. Thus

$$\mathcal{G}_j(F, \mathcal{T}) \subseteq \mathsf{Succ}_j^{\mathrm{ctrl}}(F, \mathcal{T}).$$

It remains to lower-bound $\mathbb{P}(\mathcal{G}_j(F, \mathcal{T}))$. If $\mathcal{G}_j(F, \mathcal{T})$ fails, then

$$\sum_{k=0}^{L_j(\mathcal{T})-1} \|e_{j,k}^{\mathrm{ctrl}}(F, \mathcal{T})\|_2^2 \geq \delta_{\mathrm{rd},j}^2.$$

By Markov's inequality,

$$\mathbb{P}_{\mathcal{T}}\left(\mathcal{G}_j(F, \mathcal{T})^c\right) \leq \frac{\mathbb{E}_{\mathcal{T}} \sum_{k=0}^{L_j(\mathcal{T})-1} \|e_{j,k}^{\mathrm{ctrl}}(F, \mathcal{T})\|_2^2}{\delta_{\mathrm{rd},j}^2}.$$

By the definition of the control risk,

$$\mathbb{E}_{\mathcal{T}} \sum_{k=0}^{L_j(\mathcal{T})-1} \|e_{j,k}^{\mathrm{ctrl}}(F, \mathcal{T})\|_2^2 = 2\mathcal{R}_j^{\mathrm{ctrl}}(F).$$

Combining the last three displays gives

$$\mathbb{P}_{\mathcal{T}}\left(\mathsf{Succ}_j^{\mathrm{ctrl}}(F,\mathcal{T})\right) \geq 1 - \frac{2\mathcal{R}_j^{\mathrm{ctrl}}(F)}{\delta_{\mathrm{rd},j}^2}.$$

The conservative version with $\mathcal{R}_j(F)$ follows from $\mathcal{R}_j^{\mathrm{ctrl}}(F) \leq \mathcal{R}_j(F)$. □

Proof of Corollary 5.2. Apply the estimation bounds of Section 4 to Problems I and II with confidence parameter $\eta/2$ each. By a union bound, with probability at least $1-\eta$ over the training sample,

$$\mathcal{R}_{\mathrm{I}}(\widehat{F}_{\mathrm{I},N}) \leq \widetilde{O}(\mathfrak{r}_{\mathrm{I}}), \qquad \mathcal{R}_{\mathrm{II}}(\widehat{F}_{\mathrm{II},N}) \leq \widetilde{O}(\mathfrak{r}_{\mathrm{II}}),$$

where replacing $\log(1/\eta)$ by $\log(2/\eta)$ is absorbed into the $\widetilde{O}(\cdot)$ factor. Since

$$\mathcal{R}_j^{\mathrm{ctrl}}(\widehat{F}_{j,N}) \leq \mathcal{R}_j(\widehat{F}_{j,N}),$$

Theorem 5.1 with $F = \widehat{F}_{j,N}$ gives

$$\mathbb{P}_{\mathcal{T}}\left(\mathsf{Succ}_j^{\mathrm{ctrl}}(\widehat{F}_{j,N},\mathcal{T})\right) \geq 1 - \widetilde{O}\left(\frac{\mathfrak{r}_j}{\delta_{\mathrm{rd},j}^2}\right), \qquad j \in \{\mathrm{I},\mathrm{II}\}.$$

Under the discrete-control rounding convention, $\delta_{\mathrm{rd},j}^2 = 1/4$, so this is $1 - \widetilde{O}(\mathfrak{r}_j)$. The hidden logarithmic factors are the same as in Section 4. □

E. A Separation Between Greedy Search and DFS

This appendix gives a diagnostic complete-binary-tree example separating an irreversible forward-only policy from a backtracking policy under depth-dependent ranking noise. The point is not to claim that DFS is uniformly preferable, but to isolate a basic computational distinction: a forward-only policy cannot recover from a wrong local commitment, whereas a DFS policy can spend additional test-time computation to revisit earlier states.

Search problem. Fix $H \geq 2$. Let

$$\mathcal{B}_H := \{0,1\}^{\leq H}$$

be the complete binary tree of depth H, rooted at the empty string $\emptyset$. For a node $u \in \{0,1\}^{<H}$, its two children are $u0$ and $u1$. Fix a unique accepting leaf

$$s^\star = (s_1, \ldots, s_H) \in \{0,1\}^H.$$

A leaf $v \in \{0,1\}^H$ is accepting if and only if $v = s^\star$.

For $i = 1, \ldots, H$, write

$$s^\star_{<i} := (s_1, \ldots, s_{i-1}), \qquad s^\star_{\leq i} := (s_1, \ldots, s_i),$$

with $s^\star_{<1} = \emptyset$. At the target-prefix node $s^\star_{<i}$, the correct child is $s^\star_{\leq i}$, and the wrong sibling is $s^\star_{<i}(1-s_i)$.

Let

$$r_i := H - i + 1.$$

Thus, if DFS first explores the wrong sibling at the i-th target-prefix decision, the wrong sibling subtree contains

$$2^{r_i} - 1$$

nodes.

Reward model. Fix a constant

$$0 < \bar{\zeta} < \frac{1}{2},$$

and define

$$L_H := \lceil \log_2 H \rceil.$$

We use the following depth-dependent ranking-error probabilities:

$$\zeta_i := \begin{cases} \bar{\zeta}, & r_i \le L_H, \\ \dfrac{1}{H^2 2^{r_i}}, & r_i > L_H. \end{cases} \tag{1}$$

Let

$$X_i \sim \mathrm{Ber}(\zeta_i), \qquad i = 1, \dots, H,$$

independently.

At the target-prefix node $s^\star_{<i}$, the reward ranking is defined as follows. If $X_i = 0$, then the correct child receives reward 1, and the wrong sibling receives reward 0:

$$\rho(s^\star_{<i}, s^\star_{\le i}) = 1, \qquad \rho(s^\star_{<i}, s^\star_{<i}(1 - s_i)) = 0.$$

If $X_i = 1$, the ranking is reversed:

$$\rho(s^\star_{<i}, s^\star_{\le i}) = 0, \qquad \rho(s^\star_{<i}, s^\star_{<i}(1 - s_i)) = 1.$$

At all off-target-prefix internal nodes, assign any fixed strict ranking, for example

$$\rho(u, u0) = 1, \qquad \rho(u, u1) = 0.$$

Thus all rewards are bounded in $[0, 1]$, and every local reward gap is equal to 1.

The schedule (1) is a hard-threshold coarse-to-fine ranking-noise model. The final $O(\log H)$ local rankings have constant error probability, while the earlier rankings have much smaller error probability. The formal separation below uses only this depth-dependent ranking model.

Policies and cost convention. The forward-only greedy policy starts at the root and repeatedly moves to the child with larger reward. It never backtracks. It succeeds if the leaf it reaches is $s^\star$.

The reward-ordered DFS policy is the usual walk-based DFS. It recursively explores children in decreasing reward order. When the current node has no unvisited child, the next DFS state is its parent; after returning to that parent, the policy may move to the next highest-reward unvisited child. The policy stops when it first visits $s^\star$.

We count node visits in the DFS walk, including the initial root visit, the first visit to $s^\star$, and revisits caused by backtracking. Thus, if the DFS walk moves

$$x \to y \to x \to z,$$

then both visits to x are counted.

Theorem E.1 (Forward-only greedy inference fails with high probability)**.** *For the construction above, the forward-only greedy policy succeeds with probability*

$$\mathbb{P}(\text{Greedy succeeds}) = \prod_{i=1}^{H} (1 - \zeta_i).$$

Moreover,

$$\mathbb{P}(\text{Greedy succeeds}) \le (1 - \bar{\zeta})^{L_H} \le C_{\bar{\zeta}} H^{-\gamma_{\bar{\zeta}}},$$

where

$$\gamma_{\bar{\zeta}} := \log_2 \frac{1}{1 - \bar{\zeta}} > 0$$

and $C_{\bar{\zeta}} > 0$ is a constant depending only on $\bar{\zeta}$. Hence

$$\mathbb{P}(\text{Greedy fails}) \ge 1 - C_{\bar{\zeta}} H^{-\gamma_{\bar{\zeta}}}.$$

Proof. Greedy reaches $s^\star$ if and only if every target-prefix decision is ranked correctly. Indeed, if $X_i = 1$ for the first time at level i, then greedy moves into the wrong sibling subtree $s^\star_{<i}(1 - s_i)$. Since this subtree does not contain $s^\star$, and since greedy never backtracks, the policy can no longer reach the accepting leaf.

Therefore, by independence,

$$\mathbb{P}(\text{Greedy succeeds}) = \mathbb{P}(X_1 = \cdots = X_H = 0) = \prod_{i=1}^{H}(1 - \zeta_i).$$

For the last L_H target-prefix decisions, we have $r_i \le L_H$, and hence

$$\zeta_i = \bar{\zeta}.$$

Thus

$$\mathbb{P}(\text{Greedy succeeds}) \le (1 - \bar{\zeta})^{L_H}.$$

Since $L_H = \lceil \log_2 H \rceil$,

$$(1 - \bar{\zeta})^{L_H} \le C_{\bar{\zeta}} H^{-\log_2(1/(1-\bar{\zeta}))},$$

which proves the claim. □

Theorem E.2 (DFS succeeds with small walk cost)**.** *Let $T^{\text{walk}}_{\text{DFS}}$ be the number of node visits in the reward-ordered DFS walk up to and including the first visit to $s^\star$, including revisits caused by backtracking. Then DFS finds $s^\star$ with probability one, and*

$$T^{\text{walk}}_{\text{DFS}} = H + 1 + 2\sum_{i=1}^{H} X_i(2^{r_i} - 1). \tag{2}$$

Consequently,

$$\mathbb{E}[T^{\text{walk}}_{\text{DFS}}] = O(H).$$

More explicitly,

$$\mathbb{E}[T^{\text{walk}}_{\text{DFS}}] \le H + 1 + 8\bar{\zeta}H + \frac{2}{H}.$$

Moreover, with probability at least $1 - H^{-3}$,

$$T^{\text{walk}}_{\text{DFS}} \le 9H + 1.$$

Thus DFS finds the target using $O(H)$ node visits with high probability.

Proof. Since $\mathcal{B}_H$ is finite and has a unique accepting leaf, unbudgeted DFS eventually visits $s^\star$ with probability one.

We first prove (2). The DFS walk always visits the $H + 1$ target-path nodes

$$\emptyset,\ s^\star_{\le 1}, \ldots, s^\star_{\le H}.$$

At the target-prefix node $s^\star_{<i}$, if $X_i = 0$, DFS moves to the correct child first, and hence it does not enter the wrong sibling subtree before finding $s^\star$. If $X_i = 1$, DFS first enters the wrong sibling subtree. This subtree has

$$m_i := 2^{r_i} - 1$$

nodes and does not contain $s^\star$.

Because we count the DFS walk rather than only first-time node expansions, exhausting this wrong sibling subtree and returning to $s^\star_{<i}$ contributes exactly $2m_i$ additional node visits. Indeed, the connected component consisting of the wrong sibling subtree together with the edge from $s^\star_{<i}$ to its wrong child has m_i edges. Before DFS can move to the correct child, it traverses each of these edges once downward and once upward. Each edge traversal adds one node visit to the walk. Hence the additional walk cost of this mistake is

$$2(2^{r_i} - 1).$$

The wrong sibling subtrees corresponding to different target-prefix levels are disjoint. Therefore

$$T_{\mathrm{DFS}}^{\mathrm{walk}} = H + 1 + 2\sum_{i=1}^{H} X_i(2^{r_i} - 1).$$

Taking expectations and using $r = H - i + 1$, we obtain

$$\mathbb{E}[T_{\mathrm{DFS}}^{\mathrm{walk}}] \le H + 1 + 2\sum_{r=1}^{L_H} \bar{\zeta}\, 2^r + 2 \sum_{r=L_H+1}^{H} \frac{1}{H^2 2^r}\, 2^r.$$

The first sum is bounded by

$$2\sum_{r=1}^{L_H} \bar{\zeta}\, 2^r \le 2\bar{\zeta}\, 2^{L_H+1} \le 8\bar{\zeta} H,$$

because $2^{L_H} \le 2H$. The second sum is bounded by

$$2 \sum_{r=L_H+1}^{H} \frac{1}{H^2} \le \frac{2}{H}.$$

Thus

$$\mathbb{E}[T_{\mathrm{DFS}}^{\mathrm{walk}}] \le H + 1 + 8\bar{\zeta} H + \frac{2}{H} = O(H).$$

It remains to prove the high-probability bound. Let E_{top} be the event that no ranking error occurs in the upper part of the tree, namely at levels with $r_i > L_H$. By the union bound,

$$\mathbb{P}(E_{\mathrm{top}}^c) \le \sum_{r=L_H+1}^{H} \frac{1}{H^2 2^r} \le \frac{1}{H^2 2^{L_H}} \le \frac{1}{H^3}.$$

On E_{top}, every wrong-first DFS exploration before $s^\star$ lies in the final L_H levels. Even if all of these low-level rankings are wrong, the total additional DFS-walk cost from wrong sibling subtrees is at most

$$2\sum_{r=1}^{L_H}(2^r - 1) \le 2^{L_H+2} \le 8H.$$

Therefore, on E_{top},

$$T_{\mathrm{DFS}}^{\mathrm{walk}} \le H + 1 + 8H = 9H + 1.$$

Hence

$$\mathbb{P}(T_{\mathrm{DFS}}^{\mathrm{walk}} \le 9H + 1) \ge \mathbb{P}(E_{\mathrm{top}}) \ge 1 - H^{-3}.$$

□

Discussion. The separation follows from two consequences of the depth-dependent ranking noise in (1). First, the final $O(\log H)$ local rankings have constant error probability. A forward-only policy must resolve all of these noisy decisions correctly in a single irreversible path, which gives only polynomially small success probability.

Second, on the event E_{top}, all wrong-first DFS explorations before $s^\star$ occur in the final $O(\log H)$ levels. The total walk cost of fully exploring and returning from all such wrong sibling subtrees is $O(H)$. The factor 2 in (2) is precisely the cost of walking down through a wrong subtree and then backtracking to the prefix node before trying the other child.

This example uses a complete binary tree. Thus wrong branches are not dead ends; they contain full subtrees. The distinction is between an irreversible forward-only commitment and a walk-based search process that can revisit earlier prefixes.