How Transformers Learn Compositional Reasoning on Trees: A Theory

Anonymous authors
Paper under double-blind review

Abstract

Compositional reasoning on trees is a fundamental form of structured multi-step computation: each internal node must identify its ordered children among many tokens, retrieve the appropriate operation, compose multiple intermediate values, and propagate the result upward. These branching dependencies make tree-structured Chain-of-Thought (CoT) reasoning substantially harder to analyze than chain-structured CoT, where each step follows a single dependency. We develop a theory of how transformers learn and generalize tree-structured CoT reasoning. We formulate compositional reasoning as bottom-up computation on trees and introduce a multi-head attention transformer with fixed calibrated semantics that jointly performs ordered child selection and function retrieval. Under independently injected identities and an equicorrelated token dictionary, gradient descent from zero reduces exactly to margin dynamics and converges under a positive drive floor. Token geometry governs the effective step, with explicit structural and semantic dependence in the iteration budget. An execution expansion and a training-loss barrier yield a shape-dependent generalization bound for unseen larger and deeper trees. Routing and deep-transformer experiments illustrate token-geometry effects, while language-model experiments measure conditional prediction loss on deeper reasoning traces.

1 Introduction

Transformer-based large language models (LLMs) have become a foundational architecture across a broad range of machine learning tasks (Vaswani et al., 2017), with recent advances demonstrating their remarkable ability for complex multi-step reasoning (Reynolds & McDonell, 2021). A central mechanism underlying this capability is Chain-of-Thought (CoT) reasoning (Wei et al., 2022), which enables transformers to decompose difficult problems into sequences of intermediate steps rather than producing answers directly. A growing body of empirical evidence (Zhang et al., 2023; Chen et al., 2025; Zhang et al., 2025), together with recent theoretical analyses (Li et al., 2023; Prystawski et al., 2023; Hu et al., 2024; Li et al., 2024c; Merrill & Sabharwal, 2024; Zhu et al., 2025), suggests that transformers can internally organize and manipulate intermediate representations to support structured multi-step reasoning.

Tree-structured reasoning problems provide a general abstraction for **compositional computation**, where a final answer is obtained by recursively solving interdependent subproblems arranged in a hierarchy. As illustrated in Figure 1, such compositional computation starts from the leaf nodes, each of which provides an observed value. Each internal node then applies its associated operation to the ordered values returned by its children, and the resulting intermediate value is passed upward to its parent. The computation terminates at the root, whose value represents the final answer. Such problems arise in many applications such as symbolic/circuit evaluation, program traces, proof trees, scene-graph visual reasoning, and compositional instruction following (Lööw et al., 2024; Reiss & Renieris, 2001; Choi et al., 2018; Zerroug et al., 2022; Huang et al., 2024a; Mitra et al., 2024; Hayati et al., 2025).

Such tree-structured compositional reasoning naturally aligns with CoT reasoning. A CoT trace can be viewed as an explicit sequentialization of the tree computation: it first identifies the available leaf values, then progressively resolves internal nodes whose children have already been computed, and finally reaches the root; when all nodes of a height are resolved in parallel, the number of

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

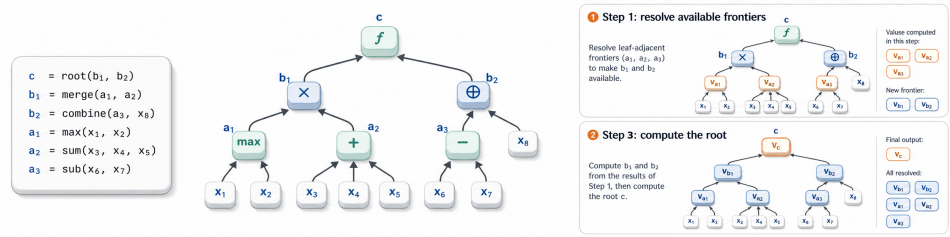


Figure 1: Tree-structured reasoning as bottom-up CoT. A program induces a tree with heterogeneous branching; each CoT step resolves the currently available frontier until the root value is obtained.

rounds equals the tree depth. To train transformer-based CoT to solve such a problem, the model must acquire several important capabilities: the model must learn to route attention to the correct children of each node, distinguish the roles of nodes, identify and apply the appropriate node-specific operation, and propagate the resulting intermediate value upward.

Recent empirical studies have demonstrated that transformers can perform CoT-style reasoning on tree-structured compositional tasks (Mitra et al., 2024; Hayati et al., 2025; Yin et al., 2025). However, the corresponding theoretical understanding remains limited[1]. The closest theoretical analyses of tree-structured composition mainly provide expressivity guarantees (Feng et al., 2023; Yehudai et al., 2025; Nandakumar et al., 2026), showing that there exist transformer weights capable of implementing CoT-style computation for tree-structured compositional reasoning. Yet these results do not explain whether practical training algorithms, such as gradient descent, can actually discover such models. This motivates our first foundational question:

(Q1) Can gradient descent provably enable transformers to acquire tree-structured compositional reasoning capabilities, such as attending to the correct children and identifying the associated compute functions?

Beyond learnability, an equally important issue is structural generalization. In practice, a trained model is compelling only if it can solve tree structures that differ from those encountered during training, such as larger trees, deeper hierarchies, or new combinations of local operations. This motivates our second question:

(Q2) Can the learned model generalize to tree structures unseen during training?

This work addresses these questions theoretically by formulating compositional reasoning as bottom-up computation on trees and introducing a multi-head attention transformer with fixed calibrated semantics that jointly performs ordered child selection and function retrieval. Because this fixed module applies the node functions, gradient descent learns the attention routing (which children to read and which function code to retrieve), not the composition functions, unlike expressivity constructions that realize both (Yehudai et al., 2025). Our main contributions are summarized as follows.

- *Optimization theory for tree-structured compositional CoT.* We prove an exact reduction of population gradient descent on the routing parameters from zero under uniform identity injection and equicorrelated tokens. A positive drive floor yields convergence of the full rollout loss. We also explicitly characterize how the conditioning of token representations affects training speed, quantifying how non-orthogonality affects optimization.
- *Structural generalization to larger and deeper trees.* We prove generalization bounds for test trees whose size and depth exceed those seen during training. The test loss is bounded through the numbers of internal outputs, execution draws, and attention candidates, given null-witness coverage and coverage of every nontrivial test head in training, shared public identities, and population training loss below a stated threshold. Combined with the convergence result, this controls test error through training accuracy.

[1]As elaborated in Section 1.1, existing theory addresses related but incomplete settings: CoT over graphs without compositional computing (Zhu et al., 2025; Xiao & Liu, 2024; Yang et al., 2025), or CoT for chain-structured compositional and state-tracking tasks (Huang et al., 2025c; Forner et al., 2026; Huang et al., 2026a). These works do not simultaneously capture structured routing and recursive composition.

- *Empirical validation on deep CoT transformers and LLMs.* We conduct experiments on attention routing models, multi-layer CoT transformers, and Qwen2.5-0.5B. The results illustrate the predicted conditioning trend beyond the exact model used in the analysis, and they measure how the loss of trained models behaves on deeper trees.

1.1 Related Works

Appendix A gives an extended discussion and a fuller list.

Compositional reasoning tasks. Empirical work motivates tree- or relation-structured compositional reasoning through tree composition, visual reasoning, vision-language evaluation, relation-based prompting, and compositional instruction or CoT training (Choi et al., 2018; Zerroug et al., 2022; Huang et al., 2024a; Mitra et al., 2024; Hayati et al., 2025; Yin et al., 2025; Li et al., 2024b). On the theory side, Feng et al. (2023); Yehudai et al. (2025) are the closest works on composition over tree-like dependencies and primarily study expressivity, as does Nandakumar et al. (2026) for bounded-depth grammars. Other compositional theories study non-tree settings: Li et al. (2023) study CoT for in-context learning of compositional functions, Huang et al. (2025c) analyze chain-structured state-tracking CoT, Wang et al. (2025) study compositional functions in deep transformers without explicit CoT, and Wies et al. (2023); Chang et al. (2025); Yao et al. (2026); Abdulsalam et al. (2026) study sub-task decomposition, controlled compositional structures, learned-skill composition, or compositional strategies built by reinforcement learning.

Theoretical studies on graph-based reasoning. Other theory studies graph-, DAG-, or path-structured reasoning, where the target is dependency propagation, navigation, or algorithmic computation rather than node-specific recursive composition. Xiao & Liu (2024) study length generalization for DAG-structured reasoning; Zhu et al. (2025) study Chain of Continuous Thought for path-finding; Yang et al. (2025) analyze CoT for symbolic path-finding on trees; Guo et al. (2025) study two-hop in-context reasoning; and Sanford et al. (2024); Abbe et al. (2024) analyze transformer computation depth and algorithmic reasoning.

Other CoT theory and training dynamics. CoT theory also studies inherently serial and hierarchical computation (Li et al., 2024c; Merrill & Sabharwal, 2024), locality and statistical models of prompting (Prystawski et al., 2023; Hu et al., 2024), sparse-dependence, parity, and regular-language tasks (Wen et al., 2025; Kim & Suzuki, 2025; Huang et al., 2025b), and in-context inference (Li et al., 2025; Huang et al., 2025a). Training-dynamics theory explains how attention specializes in in-context learning (Ahn et al., 2023; Mahankali et al., 2023; Huang et al., 2024b; Chen et al., 2024), classification (Ataee Tarzanagh et al., 2023; Vasudeva et al., 2025; Jiang et al., 2024), mixture-of-experts (Chen et al., 2022), induction heads (Nichani et al., 2024a), sparse token retrieval (Wang et al., 2024), and factual recall (Nichani et al., 2024b), all outside tree-structured CoT. Concurrent 2026 work analyzes implicit CoT (Huang et al., 2026a), CoT state-tracking dynamics (Forner et al., 2026), CoT expressivity under softmax attention and finite precision (Brösamle & Eckstein, 2026), serial depth (Brown-Cohen et al., 2026), looped latent reasoning (Fan et al., 2026), and length-generalization limits (Yang et al., 2026a;b), but not gradient-based learning of tree-structured routing.

Notations. For a vector v, $\|v\| = \|v\|_2$; for a matrix M, $\|M\| := \|M\|_F = \|\mathrm{vec}(M)\|_2$; and $x = \mathcal{O}(y)$ means $x < C|y|$ for a positive constant C.

2 Problem Setup and Transformer Architecture

We consider tree-structured compositional reasoning tasks, where computation proceeds bottom-up over a directed tree (Feng et al., 2023; Yehudai et al., 2025). As illustrated in Figure 1, leaf nodes carry observed values, and each non-leaf node applies its associated function to the values received from its children, for example $a_1 = \max(x_1, x_2)$, and passes the result upward until the root value, denoted by v_c, is computed. Nodes closer to the root correspond to higher-level subproblems.

For a general compositional reasoning task, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote its ordered rooted tree, with root $o(\mathcal{G})$ and edges $\mathcal{E} = \{(n, p(n)) : n \neq o(\mathcal{G})\}$, where $p(n)$ is the parent of n. Each leaf node $n \in \mathcal{V}$ is associated with an observed value v_n, while each non-leaf node is equipped with a function that maps the values of its children to its own value. For any node $m \in \mathcal{V}$, let $C(m)$ denote the set of

its child nodes. An internal node n has arity $a_n \le N_c$ and children $c_1(n), \dots, c_{a_n}(n)$. Values lie in a finite alphabet $[q]$, and the finite function dictionary $\mathcal{F}$ contains maps $f : [q]^{a_f} \to [q]$ of declared arity, including order-sensitive and constant functions. Leaf symbols and node functions may have any joint law. Leaves have height zero and $\mathrm{ht}(n) = 1 + \max_h \mathrm{ht}(c_h(n))$; tree depth is the root height.

2.1 Embedding Construction

For a graph $\mathcal{G}$, let $|\mathcal{V}|$ denote its number of nodes. We represent each node n by an identity token $x_n \in \mathbb{R}^d$ and its parent by $y_n = x_{p(n)}$, with zero parent and slot entries at the root. Node identities are distinct. To distinguish child roles under the same parent, we use a relation feature $r_n = \rho(y_n, l_n) \in \mathbb{R}^{d_r}$ in the spirit of attention-chooser constructions (Zhu et al., 2025), which combines the parent token with the child slot and carries immutable arity and height metadata. A node holding symbol $\xi \in [q]$ has the d-dimensional value token $v_n = \Phi e_\xi$, and a node with function $f \in \mathcal{F}$ has the function token $f_n = \Psi e_f$, where $\Phi^\top \Phi = I_q$ and $\Psi^\top \Psi = I_{|\mathcal{F}|}$. The two codes occupy separate blocks, and their null vectors are zero. Identity tokens may be nonorthogonal.

We collect all node-level tokens into $X(\mathcal{G}) = [x_n]_n$, $R(\mathcal{G}) = [r_n]_n$, $F(\mathcal{G}) = [f_n]_n$, and $U(\mathcal{G}) = [v_n]_n$. The input embedding for graph $\mathcal{G}$ is then defined as

$$\mathbf{E}(\mathcal{G}) = [X(\mathcal{G}); R(\mathcal{G}); F(\mathcal{G}); U(\mathcal{G})]. \tag{1}$$

Initially, leaves carry their observed values and null function codes; internal nodes carry function codes and null values. Only the value block changes during CoT. Let N_c denote the maximum number of child nodes associated with any non-leaf node in $\mathcal{G}$. We introduce a set of positional encoding vectors

$$\mathcal{P} = \{q_1, \dots, q_{N_c}\}, \tag{2}$$

where q_ℓ represents the ℓ-th child position under a given parent. These encodings are reused across different nodes but serve to distinguish child roles locally.

Assumption 1. *The positional encoding vectors in $\mathcal{P}$ are mutually orthonormal.*

The public identity-token set and its matrix are

$$\mathcal{B} = \{b_1, \dots, b_M\}, \qquad B = [b_1, \dots, b_M] \in \mathbb{R}^{d \times M}, \tag{3}$$

with $M \ge \max\{2, |\mathcal{V}|\}$ and an injection $\iota : \mathcal{V} \to [M]$ giving $x_n = b_{\iota(n)}$.

2.2 Transformer Architecture and Iterative CoT Process

We adopt one multi-head attention module followed by a fixed semantic module. Among $r = N_c + 1$ heads, N_c aggregate ordered child values and the remaining head retrieves the node's function. For notational simplicity, we merge the key and query matrices into a single matrix W_h^{KQ}, and fixed value projections read the appropriate blocks. The trainable parameters $\theta = (B^1, \dots, B^r)$ are separate combined bilinear score matrices, embedded in W_h^{KQ}; the remaining blocks are fixed. A similar separation has been adopted in CoT training-dynamics analysis such as Huang et al. (2025c; 2026b).

The semantic module is fixed. Given soft function coefficients c and soft operand coefficients $u_1, \dots, u_a$, each a nonnegative vector of mass at most one, it returns the distribution of the node function applied to independent draws of these inputs. Missing mass is a null symbol with zero-vector encoding. A null operand, null function code, or incompatible function arity returns null, including for constant functions. Only existing child slots are queried. The exact polynomial rule is in Appendix E; it is realized by product gates and can be approximated by smooth networks.

We now describe the iterative CoT process. At reasoning step i, let O^{i-1} denote the input to the attention layer, where $O^0 = \mathbf{E}(\mathcal{G})$ at initialization. Set $\mathcal{C}^{(h)} = \{m \ne o : \mathrm{slot}(m) = h\}$ for a child head and $\mathcal{C}^{(r)} = \mathcal{V}$ for the function head. These pools include unwritten nodes and null leaf function codes. At step i, using O^{i-1}, the eligible logits are

$$\alpha_{m,n}^{(h)} = y_m^\top B^h x_n \ (h \le N_c), \qquad \alpha_{m,n}^{(r)} = x_m^\top B^r x_n. \tag{4}$$

The positional factor acts as a hard mask: attention is zero outside $\mathcal{C}^{(h)}$, and $s^{(h)}_{m,n} = \exp(\alpha^{(h)}_{m,n}) / \sum_{j \in \mathcal{C}^{(h)}} \exp(\alpha^{(h)}_{j,n})$ inside it. Singleton reads are exact and empty pools are never queried. The output of the attention layer at step i is then obtained by concatenating the outputs of all $N_c + 1$ heads:

$$\hat{A}^i = \mathrm{Concat}\big(\mathrm{attn}_1(O^{i-1}), \ldots, \mathrm{attn}_r(O^{i-1})\big), \tag{5}$$

where $\mathrm{attn}_h(O^{i-1}) = W^V_h O^{i-1} \cdot \boldsymbol{s}^{(h)}$, and $\boldsymbol{s}^{(h)}$ denotes the corresponding attention matrix with entries $s^{(h)}_{m,n}$; child heads read value tokens and the function head reads function tokens; the fixed module uses their coefficients in the orthonormal bases. Missing child slots are inactive.

At step i, write exactly the height-i nodes and then freeze them. All reads use the previous embedding. Let $M^i = \mathrm{diag}(\mathbf{1}\{\mathrm{ht}(n) = i\})$ and let $\widetilde{Z}^i$ contain the proposed semantic outputs. Since every selected column previously had zero value, the update is

$$O^i = O^{i-1} + [0; 0; 0; \Phi \widetilde{Z}^i] M^i, \qquad O^0 = \mathbf{E}(\mathcal{G}). \tag{6}$$

After k steps, the final embedding $\hat{O}(\mathcal{G}; \theta) = O^k$ contains all predicted internal values. The rollout is free running: later steps read these predictions without replacing them by ground truth. We call this architecture and schedule the *canonical model*: fixed calibrated semantics, hard slot masks, the height schedule, and orthonormal semantic codes. Appendix E states it formally and labels these ingredients (M1) to (M4). They constrain the model, not the task, which allows any finite alphabet, any finite function dictionary of bounded arity, and any law over bounded ordered trees.

3 Expressive Capacity

We first show that, independent of optimization, the attention-MLP transformer of Section 2.2 implements the desired bottom-up CoT steps in the large-margin limit on any tree, assuming sufficient diversity among public tokens.

Assumption 2. *The public dictionary B has full column rank (hence $d \geq M$), and there exists a constant $\gamma > 0$ such that $\sigma_{\min}(\mathcal{B}) \geq \gamma$, where $\sigma_{\min}(\mathcal{B})$ denotes the minimum singular value of the matrix formed by vectors in $\mathcal{B}$.*

Assumption 2 is a standard non-degeneracy condition: node tokens are linearly independent and do not collapse. It is strictly weaker than the orthonormality imposed in prior work, which forces $\sigma_{\min}(\mathcal{B}) = 1$.

Under these assumptions, the following attention-chooser theorem shows that one multi-head attention layer can route every node to its ordered children and to its own function token for the MLP.

Theorem 1 (Attention chooser for tree-structured CoT)**.** *Suppose Assumptions 1 and 2 hold, and instantiate the relation encoder as $\rho(y, l) = l \otimes y$ with zero score weights on the arity and height metadata. For any separation scales $\{\beta_h > 0\}_{h=1}^{N_c+1}$, there exist attention matrices $\{W^V_h, W^{KQ}_h\}_{h=1}^{N_c+1}$ such that, for every graph $\mathcal{G}$ with injective public identities, every target node n, and every valid child slot $h \in [N_c]$, the head-h logits on $m \in \mathcal{C}^{(h)}$ satisfy*

$$\alpha^{(h)}_{m,n} = \begin{cases} \beta_h, & m = c_h(n), \\ 0, & m \neq c_h(n), \end{cases} \tag{7}$$

where $c_h(n)$ denotes the h-th child of node n. Moreover, the $(N_c + 1)$-th head satisfies

$$\alpha^{(N_c+1)}_{m,n} = \begin{cases} \beta_{N_c+1}, & m = n, \\ 0, & m \neq n. \end{cases} \tag{8}$$

Consequently, as $\min_h \beta_h \to \infty$, the softmax attention outputs converge to $\mathrm{attn}_h(O)_n \to v_{c_h(n)}$ for $h \in [N_c]$ and $\mathrm{attn}_{N_c+1}(O)_n \to f_n$. Outside the slot pool the attention mass is zero. Thus, in this limit, the attention layer provides the ordered child values and function token needed by the MLP for one bottom-up CoT update.

The construction places $q_h \otimes B^h$ in the relation-to-identity score block, where full column rank lets B^h interpolate $x^\top B^h x' = \beta_h \mathbf{1}\{x = x'\}$ on the public tokens; the proof is in Appendix C.

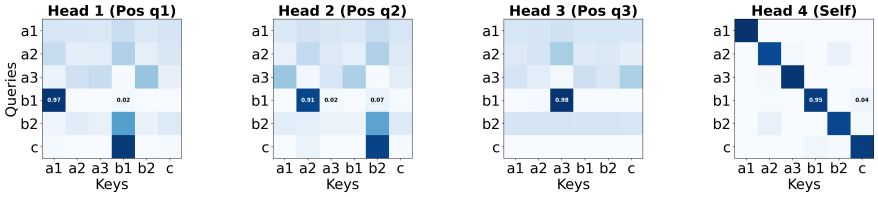


Figure 2: Routing pattern of Theorem 1 reached by directly supervised attention on a three-layer tree with nodes $\{a_1, a_2, a_3, b_1, b_2, c\}$. Heads 1 to 3 select the ordered children of the supervised queries b_1 and c, and head 4 attends to each node itself for function retrieval. Details are in Appendix B.

To illustrate the routing construction, we train the four heads on the toy tree of Figure 1 by direct cross-entropy supervision toward the target selections, not the rollout loss of Section 4. Figure 2 shows heads 1 to 3 concentrating on the appropriate ordered children and head 4 attending to each node itself, matching the limiting routing pattern in Theorem 1.

Having established exact attention routing for one bottom-up update, we next show that the resulting transformer performs iterative CoT updates on tree-structured graphs, with the number of reasoning steps equal to the depth of the tree.

Proposition 1. *Under Assumptions 1 and 2, for any graph $\mathcal{G}$ of depth k, the constructed attention-MLP transformer computes the root value, i.e., $\hat{v}_{o(\mathcal{G})} = v_{o(\mathcal{G})}$, at step k of the structural schedule in the large-margin limit.*

The proof of Proposition 1 is given in Appendix D. All nodes at the same height are updated in parallel within one CoT iteration, so a depth-k tree is solved in the large-margin limit in k reasoning steps, independent of the total number of nodes. Unlike the two-layer construction of Yehudai et al. (2025), which generates CoT tokens one node at a time, our layer aggregates children and retrieves functions for all nodes of a height in parallel, which keeps the training dynamics analyzable.

4 Training Dynamics and Convergence Analysis

Section 3 constructs attention parameters that realize the desired bottom-up CoT computation. We now study when gradient descent learns the routing mechanism from the full rollout loss, and how token conditioning controls its effective step. Specifically, we sample tree-structured reasoning tasks from a training distribution and optimize the attention parameters by gradient descent (GD).

4.1 Training Process

Graphs are drawn from a fixed distribution $\mathcal{F}(\mathcal{G})$ with depth at most K, at most $\bar{\mathcal{V}}$ nodes, and at most I_* internal nodes, through the following generation process.

Definition 1 (Graph and embedding generation)**.** *Sample an ordered tree of depth $k \in \{1, \ldots, K\}$ with arity at most N_c. Sample $|\mathcal{V}|$ distinct public tokens uniformly without replacement and assign them uniformly to nodes, independently of the shape and semantics. Assign positional encodings by child slot, and form relation features with immutable arities and heights. Sample leaf symbols and internal functions from any joint finite semantic law conditional on shape, using null function codes at leaves and null initial values at internal nodes.*

Based on the graph generation process, we adopt the training objective in terms of the discrepancy between the model outputs and the target ground truth:

$$\mathcal{L}_{\text{train}}(\theta) = \frac{1}{2}\mathbb{E}_{\mathcal{G}\sim\mathcal{F}(\mathcal{G})}\|O(\mathcal{G}) - \hat{O}(\mathcal{G};\theta)\|_F^2, \tag{9}$$

where $\hat{O}(\mathcal{G};\theta)$ denotes the final output derived by (6), and $O(\mathcal{G})$ denotes the ground truth for tree $\mathcal{G}$. The target includes every internal value. Each head starts at $B^h = 0$, giving uniform attention in each eligible pool, as is common in CoT analyses (e.g., Yang et al., 2025). Then we apply population GD:

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla_\theta\mathcal{L}_{\text{train}}(\theta^{(t)}), \qquad \theta^{(0)} = 0, \tag{10}$$

where $\eta > 0$ is the learning rate and t is the training round.

4.2 Theoretical Results

We first make the following mild assumption to bound the token norm.

Assumption 3 (Bounded tokens)**.** *Assume there exists a constant $C > 0$ such that any vector $z \in \{x_m, y_m, f_m, v_m\}_{m=1}^{|\mathcal{V}|}$ satisfies $\|z\|^2 \le C$.*

Write $Z^h = B^\top B^h B$ and $R = I_M - \mathbf{1}\mathbf{1}^\top/M$. On $Z^h = d_h R$, the correct attention mass for a pool of size N is $p(d_h; N) = e^{d_h}/(e^{d_h} + N - 1)$. Let $F(d)$ denote the actual population rollout loss on these score tables, and $\Lambda_h(d) = -\partial_{d_h} F(d)$ its corrective drive.

Theorem 2 (Exact reduction)**.** *Under the canonical model, Assumptions 1 to 3, and uniform independent identity injection, assume the equicorrelated Gram*

$$B^\top B = \lambda_\perp I_M + \beta \mathbf{1}\mathbf{1}^\top, \qquad \lambda_\perp > 0, \quad \lambda_\perp + M\beta > 0. \tag{11}$$

For every finite $\eta > 0$, GD from zero satisfies $Z_t^h = d_{h,t} R$ exactly, with

$$d_{h,t+1} = d_{h,t} + \alpha \Lambda_h(d_t), \quad d_{h,0} = 0, \qquad \alpha = \frac{\eta \lambda_\perp^2}{M-1} \ge \frac{\eta \sigma_{\min}^4(B)}{M-1}. \tag{12}$$

No equality between different heads or drive assumption is required.

The proof in Appendix G uses identity-permutation and column-shift invariance. The equicorrelated Gram allows nonorthogonal tokens and any positive minimum singular value.

Let $\mathcal{A}$ be the heads queried with a pool of size at least two on some supported shape, and $\rho = |\mathcal{A}|$. Put $D = r \sum_{j=0}^{K-1} N_c^j$, a bound on the number of draws in an unfolded execution, and let $N_* \ge 2$ bound pool sizes. For complete b-ary trees one may use $D = (b+1)I$ with I internal nodes. Define

$$\begin{aligned} L_d &= 16 I_* D^2, \qquad b_* = \log\big(2D(N_* - 1)\big), \\ g_h &= \inf\{\Lambda_h(d) : d \in [0,\infty)^r,\ d_h \le b_*\}. \end{aligned} \tag{13}$$

Theorem 3 (Convergence from zero under a drive floor)**.** *Under the hypotheses of Theorem 2, suppose* (P) *$g_h > 0$ for every $h \in \mathcal{A}$ and $0 < \alpha \le 1/L_d$. Every active margin stays nonnegative and crosses b_* permanently by*

$$T_1 = \max_{h \in \mathcal{A}} \left\lceil \frac{b_*}{\alpha g_h} \right\rceil. \tag{14}$$

Writing $F_t = F(d_t)$, for integers $s \ge 0$ and $F_{T_1} > 0$,

$$F_{T_1+s} \le \left[F_{T_1}^{-1} + \frac{\alpha(1 - \alpha L_d/2)}{\rho} s \right]^{-1}. \tag{15}$$

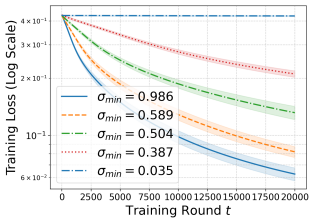
A sufficient budget for loss at most $\varepsilon > 0$ is $T_1 + \lceil \rho/(\alpha(1 - \alpha L_d/2)\varepsilon) \rceil$. Zero loss remains zero. If $\mathcal{A}$ is empty, the loss and budget are zero.

The proof is in Appendix H. The structural schedule and calibrated semantics yield a large-margin descent inequality; the drive floor supplies entry into that region.

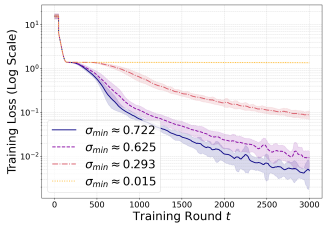
Token geometry controls the effective step. The exact step is proportional to $\lambda_\perp^2$, the squared contrast eigenvalue of the token Gram matrix. Since $\lambda_\perp \ge \sigma_{\min}^2(B)$, with $g_{\min} = \min_{h \in \mathcal{A}} g_h$ the budget is at most

$$O\left(1 + \frac{M}{\eta \sigma_{\min}^4(B)} \left[\frac{1 + \log D + \log N_*}{g_{\min}} + \frac{\rho}{\varepsilon} \right] \right).$$

For complete b-ary trees, $D = \Theta(V)$, $I = \Theta(V/b)$, and $L_d = \Theta(V^3/b)$, giving polynomial structural factors in graph size. If $\beta \ge 0$, then $\lambda_\perp = \sigma_{\min}^2(B)$ and the conditioning dependence is exact. If $\beta < 0$, the minimum eigenvalue can instead lie in the common-token direction, which does not drive routing. The bound in terms of $\sigma_{\min}$ remains valid, but a larger $\sigma_{\min}$ alone need not imply faster actual training across such dictionaries. Comparisons hold at a fixed admissible η and fixed task law.



(a) Supervised routing heads.



(b) Jointly trained deep transformer.

Figure 3: Training loss under different token conditionings. Tokens have unit norm, and in these sweeps a smaller $\sigma_{\min}(\mathcal{B})$ comes with a larger $\sigma_{\max}(\mathcal{B})$. (a) Routing heads with per-head supervision. (b) A three-layer attention network trained end to end. Mean and one standard deviation over five runs in (a) and ten runs in (b).

Remark 1. *The box version of Condition (D) (Definition 2) requires $-\partial_{\gamma_h} F_\sigma(\gamma) \geq c(1-\gamma_h)$ for some $c > 0$, every supported shape σ, $h \in \mathcal{A}$, and $\gamma \in [0,1]^{|\mathcal{A}|}$, where $F_\sigma(\gamma)$ is the loss on σ when each head h reads its correct source with weight γ_h and the uniform pool average otherwise. If every active head has at least two candidates on every supported shape, it gives $g_h \geq c/(4\bar{\mathcal{V}}D^2(N_*-1)^2)$ (Corollary 1); height-two families provide certified illustrations (Appendix K). For arbitrary full-rank dictionaries, Proposition 3 and Theorem 8 give stationarity and a local zero-loss basin with rate proportional to $\sigma^4_{\min}(B)$, but no from-zero global convergence.*

Numerical validation. In Figure 3, a larger measured $\sigma_{\min}(\mathcal{B})$ yields faster loss decay. Unit-norm tokens rule out uniform rescaling as the cause, although they do not hold $\sigma_{\max}(\mathcal{B})$ or the logits fixed; a separate routing control with $\sigma_{\max}(\mathcal{B}) = 1$ fixed shows the same ordering. Panel (a) supervises head contents directly and panel (b) trains end to end, so both are qualitative illustrations beyond the calibrated model; see Appendix B.

5 Generalization Analysis

We next study whether the learned model generalizes to deeper and larger graphs unseen during training. The challenge is error propagation: one-step routing errors are reused in later CoT steps and accumulate toward the root. At test time we apply the same canonical update for the height of the test tree, with injective identities from the same public token set. The graph-level test loss of $\mathcal{G}'$ is defined as

$$\mathcal{L}_{\text{test}}(\mathcal{G}';\theta) = \frac{1}{2}\|O(\mathcal{G}') - \widehat{O}(\mathcal{G}';\theta)\|_F^2, \tag{16}$$

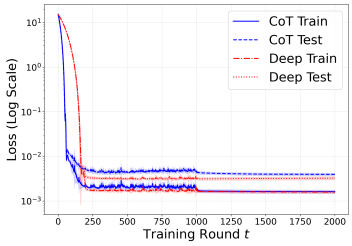
where $O(\mathcal{G}')$ is the ground-truth vector for graph $\mathcal{G}'$. The following result transfers small population training loss to every compatible test instance. A head is structurally nontrivial if it is queried with a nonsingleton pool on a supported training shape. A null witness is an internal target and a wrong candidate that is necessarily null when read: an unwritten node for a child head, or a leaf for the function head. Let W_h^{wit} be its expected count for head h, and $W_{\min}$ the minimum over structurally nontrivial training heads.

Theorem 4 (Generalization from structural witnesses)**.** *Use the canonical model, any bounded training law of shapes and semantics, and uniform independent injection into the public set of M identities. Assume the set $\mathcal{A}$ of structurally nontrivial training heads is nonempty, $W_{\min} > 0$, and their pool sizes are at most $N_* \geq 2$. For any finite attention parameters with $\mathcal{L}_{\text{train}}(\theta) \leq \varepsilon$, suppose*

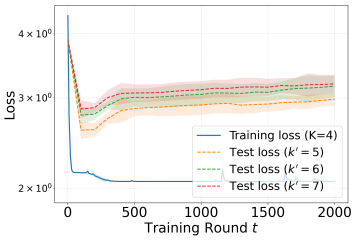
$$0 < \varepsilon \leq \varepsilon_{\text{gen}} := \frac{W_{\min}}{8M(M-1)(N_*-1)^2}. \tag{17}$$

Let $\mathcal{G}'$ have I' internal nodes and V' nodes in total. Write $N'_h = |\mathcal{C}^{(h)}|$ for its child-pool sizes and $\mathcal{I}'(n)$ for the internal nodes in the true subtree rooted at n. Define

$$A_n = \sum_{v \in \mathcal{I}'(n)} \left[V' - 1 + \sum_{h=1}^{a_v} (N'_h - 1) \right], \qquad \Gamma(\mathcal{G}') = \sum_{n \in \mathcal{I}'} A_n^2.$$



(a) CoT block versus six-layer stack.



(b) LLM loss on deeper trees.

Figure 4: Empirical tests beyond the canonical assumptions. (a) A weight-tied CoT block versus an untied non-CoT six-layer stack, both supervised only at the root, on node-budgeted trees. (b) Qwen2.5-0.5B (Qwen Team, 2024) fine-tuned at depth $K = 4$ and evaluated at depths $k' \in \{5, 6, 7\}$ by teacher-forced completion loss. Mean and one standard deviation over five seeds.

Assume M bounds both training and test node counts, test identities are injective, and every head nontrivial on $\mathcal{G}'$ belongs to $\mathcal{A}$. Then, for every assignment of test functions and leaf values from the same dictionaries,

$$\mathcal{L}_{\text{test}}(\mathcal{G}'; \theta) \le \min\left\{ I', \frac{8\Gamma(\mathcal{G}')M(M-1)}{W_{\min}}\varepsilon \right\}. \tag{18}$$

The proof in Appendix L bounds the first wrong draw in each true subtree and uses the barrier of Corollary 2. No symmetry or rank condition on the identity Gram matrix is required. For complete trees of positive height and arity $b \ge 2$, $\Gamma = \Theta(V'^4/b^2)$, giving a coefficient $O(V'^4M^2/(b^2W_{\min}))$. The appendix gives the counting argument, a root-answer bound, and a training budget from Theorem 3.

Depth and capacity. For complete trees with fixed $b \ge 2$, $V' = \Theta(b^{K'})$, so the sufficient training tolerance for a fixed test-loss target decreases exponentially with test depth. This is a limitation of the bound, not a necessity result. Distinct public identities require $M \ge V'$. The full-rank optimization results further require $d \ge M$, which caps their test size at d. Neither the transfer proof nor existence of large-margin routing requires full rank (Appendix L).

Figure 4 presents two empirical tests. In Figure 4(a), one CoT block trained with six scheduled steps runs up to twelve steps at test time and reaches test loss comparable to, though slightly above, that of the six-layer stack; the node budget keeps the realized height at four, so this panel compares architectures rather than extrapolation in height. Figure 4(b) fine-tunes Qwen2.5-0.5B on full binary trees of depth $K = 4$, whose trace serializes the height schedule, and evaluates on depths $k' \in \{5, 6, 7\}$. The teacher-forced test loss increases monotonically with depth; it falls early in training, then rises slowly while staying above the training loss. These observations illustrate behavior on deeper trees; they measure neither free-running accuracy nor the bound's size exponent. Details are in Appendix B.

6 Conclusion and Limitations

We developed a theory of how transformers learn and generalize tree-structured compositional CoT, isolating the routing problem of ordered child selection and function retrieval. For fixed calibrated semantics and equicorrelated identities, population GD from zero reduces exactly to margin dynamics and converges under a positive drive floor, with token conditioning governing the effective step; an execution expansion and training-loss barrier give generalization bounds for larger and deeper trees. Experiments illustrate the conditioning trend and measure teacher-forced loss on deeper trees. Our analysis relies on controlled tree distributions, a fixed calibrated semantic module, a topology supplied through parent identities and relation features, and, for optimization, linearly independent tokens; identity tokens cap the test size, and for complete trees the certified training-loss tolerance shrinks exponentially with test depth. Extending from-zero convergence to arbitrary full-rank dictionaries, and the theory to learned topology, sequential traces, end-to-end multi-layer training, partial CoT supervision, and richer graph tasks, remains an important direction.

AI USE STATEMENT

In this work, we used generative AI agents, specifically Claude and Codex, to help develop the revised theoretical model and assumptions, formulate the corrected mathematical claims, provide ingredients for and assist in writing the proofs, and write and run verification computations. We also used these agents to implement and debug experiment code, draft and edit parts of the manuscript, and format references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions, canonical model, training objective, and initialization are stated in Sections 2 to 4. The appendix provides proofs, including the execution expansion, convergence analysis, and structural generalization argument. Appendix B specifies the experiment settings, plotted metrics, and run counts. Supplementary scripts implement the synthetic and language-model experiments, the routing control of Table 2, literal-rollout checks of the generalization inequalities, and exact coefficient certificates. The supplementary README distinguishes available run records from code-only artifacts. Numerical sanity checks test implementations and finite instances, not the universal statements of the theorems.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. *Advances in Neural Information Processing Systems*, 37:27850–27895, 2024.

Azwar Abdulsalam, Nishil Patel, and Andrew Saxe. RL post-training builds compositional reasoning strategies. *arXiv preprint arXiv:2607.07646*, 2026.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023.

Moritz Brösamle and Stephan Eckstein. The expressive power of low precision softmax transformers with (summarized) chain-of-thought. *arXiv preprint arXiv:2605.18079*, 2026.

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

Hoyeon Chang, Jinho Park, Hanseul Cho, Sohee Yang, Miyoung Ko, Hyeonbin Hwang, Seungpil Won, Dohaeng Lee, Youbin Ahn, and Minjoon Seo. Characterizing pattern matching and its limits on compositional task structures. *arXiv preprint arXiv:2505.20278*, 2025.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Conference on Learning Theory*, 2024.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062, 2022.

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. Learning to compose task-specific tree structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Alicia Curth, Rachel Lawrence, Sushrut Karmalkar, and Niranjani Prasad. Do transformers use their depth adaptively? evidence from a relational reasoning task. *arXiv preprint arXiv:2604.12426*, 2026.

Ying Fan, Anej Svete, and Kangwook Lee. Bridging the gap between latent and explicit reasoning with looped transformers. *arXiv preprint arXiv:2606.31779*, 2026.

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

Niklas Forner, Marcel Kühn, Matthias Thamm, and Bernd Rosenow. Learning dynamics of chain-of-thought state tracking in a solvable transformer model. *arXiv preprint arXiv:2606.18164*, 2026.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I Jordan, and Stuart Russell. How do LLMs perform two-hop reasoning in context? *arXiv preprint arXiv:2502.13913*, 2025.

Shirley Anugrah Hayati, Taehee Jung, Tristan Bodding-Long, Sudipta Kar, Abhinav Sethy, Joo-Kyung Kim, and Dongyeop Kang. Chain-of-instructions: Compositional instruction tuning on large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24005–24013, 2025.

Xinyang Hu, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Unveiling the statistical foundations of chain-of-thought prompting methods. *arXiv preprint arXiv:2408.14511*, 2024.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob Hansen, Sivan Doveh, Victor Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, et al. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946, 2024a.

Jianhao Huang, Zixuan Wang, and Jason D Lee. Transformers learn to implement multi-step gradient descent with chain of thought. In *International Conference on Learning Representations*, 2025a.

Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *International Conference on Machine Learning*, 2025b.

Yixiao Huang, Hanlin Zhu, Zixuan Wang, Jiantao Jiao, Stuart Russell, Somayeh Sojoudi, and Song Mei. Transformers provably learn to internalize chain-of-thought. *arXiv preprint arXiv:2605.28600*, 2026a.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *International Conference on Machine Learning*, pp. 19660–19722. PMLR, 2024b.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. In *Advances in Neural Information Processing Systems*, 2025c.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the learning dynamics of rlvr at the edge of competence. *arXiv preprint arXiv:2602.14872*, 2026b.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. *Advances in Neural Information Processing Systems*, 37:135464–135625, 2024.

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

Hongkang Li, Songtao Lu, Pin-Yu Chen, Xiaodong Cui, and Meng Wang. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. In *International Conference on Learning Representations*, 2025.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36:22021–22046, 2023.

Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024a.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. Understanding and patching compositional reasoning in llms. *arXiv preprint arXiv:2402.14328*, 2024b.

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

Andreas Lööw, Daniele Nantes-Sobrinho, Sacha-Élie Ayoun, Caroline Cronjäger, Petar Maksimović, and Philippa Gardner. Compositional symbolic execution for correctness and incorrectness reasoning (extended version). *arXiv preprint arXiv:2407.10838*, 2024.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

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

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

Vinoth Nandakumar, Qiang Qu, Pramod Thebe, Sakshi Khachariya, and Tongliang Liu. An expressivity analysis of hierarchical modelling in deep transformers via bounded-depth grammars. *arXiv preprint arXiv:2606.17522*, 2026.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070, 2024a.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. *arXiv preprint arXiv:2412.06538*, 2024b.

Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36:70926–70947, 2023.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Steven P Reiss and Manos Renieris. Encoding program executions. In *Proceedings of the 23rd International Conference on Software Engineering. ICSE 2001*, pp. 221–230. IEEE, 2001.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–7, 2021.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *International Conference on Machine Learning*, pp. 43276–43327. PMLR, 2024.

Christos Thrampoulidis. Implicit optimization bias of next-token prediction in linear models. *Advances in Neural Information Processing Systems*, 37:22624–22656, 2024.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 51854–51912, 2024.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *International Conference on Learning Representations*, 2025.

Noam Wies, Yoav Levine, and Amnon Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. In *International Conference on Learning Representations*, 2023.

Changnan Xiao and Bing Liu. A theory for length generalization in learning to reason. *arXiv preprint arXiv:2404.00560*, 2024.

Andy Yang, Pascal Bergsträßer, Georg Zetzsche, David Chiang, and Anthony W. Lin. Length generalization bounds for transformers. *arXiv preprint arXiv:2603.02238*, 2026a.

Andy Yang, Blerta Veseli, Corentin Barloy, Michaël Cadilhac, Andreas Krebs, Charles Paperman, Howard Straubing, and Michael Hahn. Algebraic decomposition theory for transformer length generalization. *arXiv preprint arXiv:2608.13433*, 2026b.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. In *Advances in Neural Information Processing Systems*, 2025.

Xinhao Yao, Ruifeng Ren, Yun Liao, Lizhong Ding, and Yong Liu. Compositional generalization from learned skills via CoT training: A theoretical and structural analysis for reasoning. In *International Conference on Learning Representations*, 2026.

Gilad Yehudai, Noah Amsel, and Joan Bruna. Compositional reasoning with transformers, rnns, and chain of thought. *arXiv preprint arXiv:2503.01544*, 2025.

Fangcong Yin, Zeyu Leo Liu, Liu Leqi, Xi Ye, and Greg Durrett. Learning composable chains-of-thought. *arXiv preprint arXiv:2505.22635*, 2025.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35: 29776–29788, 2022.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *arXiv preprint arXiv:2505.12514*, 2025.

A Additional Related Work

Compositional reasoning tasks. Empirical work has studied compositional reasoning in several structured domains. Choi et al. (2018) study learning task-specific tree structures, providing an example of tree-structured composition in neural models. Zerroug et al. (2022) introduce a benchmark for compositional visual reasoning, and Huang et al. (2024a) study compositional reasoning evaluation for modern vision-language models. Mitra et al. (2024) propose compositional CoT prompting for multimodal reasoning, Hayati et al. (2025) study compositional instruction tuning, Yin et al. (2025) study learning composable chains of thought, and Li et al. (2024b) investigate compositional reasoning behavior in LLMs.

On the theory side, Feng et al. (2023) and Yehudai et al. (2025) are closest to our setting because they consider multi-step reasoning tasks with an underlying tree-like compositional dependency and study how transformers, with or without CoT, can implement the corresponding computation. Feng et al. (2023) give a theoretical perspective on CoT for recursive reasoning problems and study how CoT can help transformers realize multi-step computations. Yehudai et al. (2025) study Compositional Reasoning Questions, including Boolean formulas and multi-step word problems, and compare transformers, RNNs, and transformers augmented with CoT tokens. For example, Yehudai et al. (2025) consider tasks where each internal node selects an answer from its children via an $\arg\max$ over inner products. Their transformer employs hardmax attention, so each token attends to a single input, enabling clean expressivity constructions. To solve compositional tasks with CoT, they design a two-layer transformer that sequentially generates CoT tokens: the first layer solves local sub-questions, while the second enforces traversal by routing the last CoT token to the next node. Nandakumar et al. (2026) analyze how deep transformers represent hierarchical structure through bounded-depth context-free grammars, constructing positional-attention transformers whose depth grows linearly with grammar depth. Abdulsalam et al. (2026) ask whether reinforcement-learning post-training composes primitive skills into higher-level strategies in a controlled rewrite-grammar environment, and report a phased compositional mechanism. Motivated by this line of expressivity analysis, our architecture implements CoT as a layer-wise bottom-up update with a fixed calibrated semantic module, where all nodes at the same height are updated in parallel, enabling a direct training-dynamics analysis of the attention-routing mechanism.

Several theoretical works study compositional tasks outside tree-structured recursive composition. Li et al. (2023) study the effect of CoT on in-context learning of compositional functions, modeled by multi-layer perceptrons, and identify filtering and single-step learning as two components of the CoT mechanism in that setting. Huang et al. (2025c) analyze gradient-based training of one-layer transformers on synthetic state-tracking (LEGO) tasks with explicit CoT supervision; their theory characterizes how algebraic structure governs attention concentration and length generalization, and further establishes recursive self-training guarantees for extending solvable reasoning length. Wang et al. (2025) study the learnability of compositional functions under deep transformer architectures, showing that gradient-based training can recover k-fold permutation compositions when equipped with easy-to-hard curriculum learning. Wies et al. (2023) prove that adding sub-task supervision can make certain sequence-to-sequence composite tasks learnable. Chang et al. (2025) analyze pattern matching and sample complexity in controlled compositional task structures, and Yao et al. (2026) study compositional generalization from learned skills under CoT training.

Graph-based and algorithmic reasoning. A separate line of theory studies reasoning over graphs, DAGs, paths, or algorithmic structures. Xiao & Liu (2024) develop length-generalization theory for reasoning processes represented as DAGs, with applications to algorithmic reasoning tasks such as parity and arithmetic. Zhu et al. (2025) study Chain of Continuous Thought for path-finding, where continuous latent reasoning states are used to represent intermediate path computations. Yang et al. (2025) study gradient-based learning of one-layer multi-head transformers for symbolic path-finding on trees via CoT, considering both backward reasoning (goal-to-root) and forward reasoning (root-to-goal). They show that gradient descent trains different attention heads to specialize and coordinate, enabling a two-stage reasoning process within a single autoregressive CoT trace and yielding generalization guarantees to unseen trees. Guo et al. (2025) analyze how LLMs perform two-hop reasoning in context. Sanford et al. (2024) relate transformer depth to parallel computation and logarithmic-depth algorithms, and Abbe et al. (2024) study algorithmic reasoning through the globality barrier and inductive scratchpads. More recently, Yang et al. (2026a) show that computable length-generalization bounds do not exist for two-layer C-RASP, and hence for transformers, while

giving a computable bound for a positive fragment, and Yang et al. (2026b) characterize exactly which regular languages transformers length-generalize on, with a polynomial-time decision algorithm. Curth et al. (2026) ask empirically whether transformers use their depth adaptively on a multi-hop relational reasoning task, using early readouts and causal patching.

Other theoretical studies of CoT. Beyond compositional and graph-based tasks, CoT theory has studied several other formal settings. Li et al. (2024c) analyze how CoT enables transformers to solve inherently serial problems, and Merrill & Sabharwal (2024) study the expressive power of transformers augmented with intermediate reasoning. Prystawski et al. (2023) study step-by-step reasoning from the perspective of locality of experience, and Hu et al. (2024) analyze statistical foundations of CoT prompting. Wen et al. (2025) study sample-efficiency gains from CoT in sparse-dependence tasks, Kim & Suzuki (2025) study parity with CoT, Li et al. (2025); Huang et al. (2025a) study CoT for in-context weight prediction, and Huang et al. (2025b) study training dynamics for regular language recognition. In particular, Wen et al. (2025) study how CoT improves sample efficiency for parity learning by transforming sparse input dependencies into sparse attention patterns, showing a polynomial-versus-exponential sample-complexity separation between CoT and non-CoT learning. Huang et al. (2025b) analyze one-layer transformers on regular language recognition problems such as even-pairs and parity-check sequence classification, showing a two-phase training behavior where attention first separates inputs and the linear layer subsequently converges via implicit bias. Li et al. (2025) and Huang et al. (2025a) analyze transformers trained with CoT on in-context weight prediction, showing how CoT enables one-layer models to implement multi-step update procedures. These works provide theoretical models of intermediate reasoning in settings different from tree-structured recursive composition. Concurrent 2026 studies extend several of these directions. Huang et al. (2026a) give a theoretical analysis of implicit CoT, proving that an L-layer transformer trained under a staged curriculum learns k-parity with polynomially many samples while avoiding explicit reasoning tokens at inference. Forner et al. (2026) derive mean-field learning dynamics for a solvable one-block transformer trained on state sequences generated by composing permutations, separating attention retrieval from a logic module that applies the retrieved permutation. Brösamle & Eckstein (2026) establish CoT expressivity for standard softmax decoders with rounded activations and attention weights, avoiding the hardmax attention and unbounded precision of earlier constructions. Brown-Cohen et al. (2026) formalize opaque serial depth, the longest computation that can be carried out without interpretable intermediate steps, and compute upper bounds for current models. Fan et al. (2026) show that looped, recurrent-depth transformers with supervision at each latent position can close the gap between latent and explicit CoT. These settings are chain-structured, language-theoretic, or latent; none analyzes gradient-based learning of ordered child selection and function retrieval on trees, which is the subject of our optimization and generalization results.

Training dynamics of transformers for other problems. Separate from CoT, a large training-dynamics literature analyzes how transformer components specialize under gradient-based optimization. For in-context learning, Ahn et al. (2023); Mahankali et al. (2023); Huang et al. (2024b); Chen et al. (2024) study how attention-based models implement or converge to regression-like update rules. Chen et al. (2022) study training dynamics in mixture-of-experts models. Ataee Tarzanagh et al. (2023); Vasudeva et al. (2025); Jiang et al. (2024) analyze classification and implicit-bias phenomena in attention models. For language-modeling mechanisms, Tian et al. (2023); Li et al. (2024a); Thrampoulidis (2024) study next-token prediction, Nichani et al. (2024a) study induction-head formation, Wang et al. (2024) study sparse token retrieval, and Nichani et al. (2024b) analyze factual recall through associative-memory mechanisms.

B Experiment Details

Overview. Table 1 lists, for each figure, the trained model, its supervision, the plotted quantity, and the number of independent runs; curves from several runs show the mean and one standard deviation. We report losses because our bounds are stated for losses. Exact-match accuracy of free-running generation is a natural complement for the Boolean tasks and is not included in these figures.

Compute resources. The LLM experiments were run on a compute server with an NVIDIA H200 NVL GPU (140 GB HBM3e), two AMD EPYC 9555 processors (128 physical cores in total), and

Table 1: Experiment summary. A run is an independent draw of data and initialization.

Figure	Model	Supervision	Plotted quantity	Runs
2	Four attention heads	Cross-entropy toward target selections	Attention weights	1
3(a)	Four attention heads	Per-head squared error toward target contents	Training loss	5
3(b)	Three-layer attention network	Squared error at every updated node	Training loss	10
4(a)	CoT block and six-block stack	Squared error at the root only	Root loss, training and test	5
4(b)	Qwen2.5-0.5B with LoRA	Token cross-entropy on the completion	Teacher-forced completion loss	5
Table 2	One routing head	Squared routing error	Steps to loss 0.01	5

approximately 2.2 TB of system RAM (CUDA 13.0, driver 580.126.20). The synthetic experiments are small enough to run on a CPU.

Figure 2: We visualize the converged softmax attention patterns on a three-layer tree with six nodes $\{a_1, a_2, a_3, b_1, b_2, c\}$. Each node is assigned an 8-dimensional embedding sampled from a standard Gaussian and normalized to unit norm. We use four independent attention heads: three positional heads corresponding to potential child slots and one self-attention head. Each head is parameterized by a bilinear matrix B^h with independent Gaussian entries of standard deviation 0.01, and every node is a candidate key during training.

Training is performed directly on the attention parameters, not through the rollout loss of Section 4, using full-batch gradient descent with learning rate $\eta = 0.05$ for 1000 epochs. For each head, we specify supervision targets corresponding to the desired routing behavior: heads 1 to 3 are trained to attend from the queries b_1 and c to their children in the corresponding slot ($b_1 \to a_1, a_2, a_3$ and $c \to b_1, b_2$), while head 4 is trained to attend from each node to itself. Optimization minimizes the cross-entropy between the predicted attention distributions and these target selections. The run uses a fixed random seed.

After training, we compute the full 6×6 attention matrices $\mathrm{softmax}(XB^hX^\top)$ for each head and visualize them as heat maps. Queries without a target in a head, such as the leaves in heads 1 to 3, receive no training signal in that head. For display only, the row for query c in head 3 is set to zero after training, since c has two children; the other unsupervised rows remain unmasked. The resulting plots illustrate that each head concentrates its mass on the intended tokens for the supervised queries while suppressing irrelevant entries, matching the extreme routing regime characterized in Theorem 1.

Figure 3(a): This experiment isolates the effect of token conditioning on routing. Each tree has three levels: the root and each of its children receive between one and three children uniformly at random, and the third level consists of leaves, so a tree has at most 13 nodes. Leaves carry random bits encoded by fixed 12-dimensional sinusoidal vectors, and every internal node carries one of four gate codes (majority, conjunction, disjunction, parity), encoded one-hot. For each tree, the node tokens are

$$t_n = \frac{(1-\lambda)q_n + \lambda u}{\|(1-\lambda)q_n + \lambda u\|},$$

where the q_n are rows of a random orthogonal matrix in $\mathbb{R}^{12\times 12}$ (Gaussian rows for the rare trees with 13 nodes) and $u = \mathbf{1}/\sqrt{12}$ is a common direction. The mixing weight $\lambda \in \{0, 0.4, 0.5, 0.6, 0.95\}$ sets $\sigma_{\min}(\mathcal{B})$, and the legend reports its average over trees and seeds. Every token has unit norm, so a tree with n nodes has a token matrix of Frobenius norm $\sqrt{n}$. In these mixing sweeps, $\sigma_{\max}(\mathcal{B})$ grows from about 1.0 to about 2.6 as $\sigma_{\min}(\mathcal{B})$ decreases. This rules out uniform rescaling of the tokens as the explanation, but it does not fix the logits or isolate the minimum singular value. Three positional heads and one function head score every node of the tree by $x_n^\top B^h x_m$, with B^h initialized with independent Gaussian entries of standard deviation 0.05. The target of a positional head is the value vector of the child in its slot, and the target of the function head is the node's own function code. Internal value vectors remain zero, so this diagnostic trains routing without evaluating recursive composition, and no semantic module is used. The plotted loss

is half the summed squared error between attended contents and targets, divided by the number of supervised node and head pairs; each update divides the summed gradient by the number of trees and uses learning rate 0.05. We run 2×10^4 full-batch updates on 64 trees and report the mean and one standard deviation over five seeds, each drawing fresh trees, tokens, and initializations.

Control with fixed $\sigma_{\max}(\mathcal{B})$. To control the maximum singular value in a separate routing task, we also train a single softmax routing head on $n = 12$ tokens, the columns of $X = U\,\mathrm{diag}(s)V^\top \in \mathbb{R}^{12\times 12}$, where U and V are random orthogonal matrices and the singular values s are evenly spaced from 1 down to $\sigma_{\min}(\mathcal{B})$, so that $\sigma_{\max}(\mathcal{B}) = 1$ in every setting. Each token must route to itself: the routing matrix P applies a softmax to each column of $X^\top W X$, the loss is $\|P - I\|_F^2/(2n)$, and W starts at zero and follows full-batch gradient descent with learning rate 8. Table 2 reports the number of steps needed to reach loss 0.01 over five seeds. All 25 runs converged, and the ordering matches Figure 3.

Table 2: Steps to reach routing loss 0.01 with $\sigma_{\max}(\mathcal{B}) = 1$ fixed, over five seeds.

$\sigma_{\min}(\mathcal{B})$	0.30	0.45	0.60	0.80	1.00
Mean steps	542.4	326.6	207.6	122.8	79.0
Range	446 to 630	292 to 356	194 to 218	120 to 125	79

Figure 3(b): Trees are generated top-down from a root at level three, with at most three children per node, until a budget of 30 nodes is reached. Nodes with children apply the mean, the coordinatewise maximum, or the coordinatewise minimum of their children's value vectors, chosen uniformly and encoded one-hot, and the 17 level-zero leaves carry random bits with the sinusoidal encoding above. The budget leaves three childless level-one records with zero targets and random function codes; they are included in the level-one update and loss. Tokens follow the construction of Figure 3(a) in dimension 12 with $\lambda \in \{0.05, 0.4, 0.7, 0.98\}$. Since trees have more nodes than dimensions, the unmixed rows are normalized Gaussian vectors, and $\sigma_{\max}(\mathcal{B})$ grows from about 2.3 to about 5.5 as $\sigma_{\min}(\mathcal{B})$ decreases. The model stacks three attention layers. Each layer has three child heads and one function head with bilinear scores over all nodes, concatenates their outputs, and applies a trainable linear output projection, a residual connection, and layer normalization; it has no feed-forward sublayer. At each CoT step, the stack is applied to the current value embeddings and the nodes of the current level take its output. All parameters are trained jointly with Adam (learning rate 5×10^{-3}, gradient clipping at one) for 3000 full-batch rounds on 32 trees, and the loss is the squared error at every node updated at every step. Curves show the mean and one standard deviation over ten runs with fresh data and initialization, smoothed by a moving average of width 50. In both panels, the legend reports the average smallest singular value from the compact singular value decomposition of each tree's token matrix, and each tree has its own tokens rather than a common public dictionary. For the 13-node trees in Figure 3(a) and for all trees here, the number of tokens exceeds the dimension 12, so the token Gram matrix is singular; the reported value is an empirical proxy for conditioning, not the full-column-rank parameter of Assumption 2. The dependence on $\sigma_{\min}(\mathcal{B})$ is qualitatively the same as in Figure 3(a), which suggests that the representation-dependent behavior characterized by our theory remains relevant for jointly trained multi-layer models.

Figure 4(a): Trees are generated top-down. The root receives the depth label k, and a node at level $\ell \geq 2$ receives $\min\{3, \max(1, \lfloor R/\ell \rfloor)\}$ children, where R is the remaining node budget (a node at level one would receive between one and three children uniformly at random), until the budget of 50 nodes is exhausted; nodes left without children become leaves. Each node draws its token uniformly, with replacement, from a pool of 50 unit-norm Gaussian vectors in $\mathbb{R}^{16}$. Internal nodes average the value vectors of their children, and each leaf value is its token scaled by a factor drawn uniformly from $[0.3, 0.7]$. Because the budget is spent near the root, every generated tree has exactly 50 nodes and realized height four, both for the training label $K = 6$ and for the test labels $k' \in \{9, 10, 11, 12\}$. The leaves sit at levels $k - 4$ and $k - 3$, and the fraction of internal nodes with a single child grows from 0.25 at $K = 6$ to 0.60 at $k' = 12$. The depth label therefore sets the length of the CoT schedule rather than the realized height.

The CoT model is a single block with three attention heads, whose scores compare the token of the query node with the parent token of each candidate, followed by averaging of the head outputs, a residual connection with layer normalization, and a feed-forward network ($16 \to 64 \to 16$, ReLU)

with a residual connection and layer normalization. It is applied for k steps; at step ℓ, every node with level ℓ takes the block output. The non-CoT baseline stacks six untied blocks of the same form and updates every node at every block, so its depth exceeds the realized height. Both models receive only the squared error at the root, and neither receives intermediate supervision. They are trained with Adam (weight decay 10^{-4}, gradient clipping at one) for 2000 full-batch rounds on 100 training trees, with learning rates 0.01 for the CoT model and 0.003 for the baseline, both decayed at rounds 1000 and 1500. The test curve averages the root loss over 20 test trees per depth and over the four test depths. Curves show the mean and one standard deviation over five seeds. This experiment compares the two architectures under shifts in shape and schedule length; it does not test extrapolation in realized height, which would require larger node budgets.

Figure 4(b): We evaluate depth generalization using Qwen2.5-0.5B (Qwen Team, 2024) on synthetic tree-structured reasoning traces. Each example is a full binary tree whose leaves carry Boolean values and whose internal nodes apply one of five Boolean functions $\{\mathsf{AND}, \mathsf{OR}, \mathsf{XOR}, \mathsf{IMP}, \mathsf{NIMP}\}$. The prompt lists node definitions using special tokens <TASK>, <DEF>, <ROOT>, and <COT>; the completion is the bottom-up CoT trace, written as <STEP> node <VAL> value for each internal node, followed by <ANS> and the root value. The completion lists the internal nodes in height order: all height-one nodes, then all height-two nodes, and so on. We train on depth-$K = 4$ trees and evaluate on unseen depths $k' \in \{5, 6, 7\}$.

We add 512 node tokens together with control, value, and function tokens to the tokenizer. Node tokens are sampled without replacement from this pool for each tree, and prompt definitions are listed in post-order. Value and function tokens are initialized from the corresponding pretrained textual embeddings. The node input embeddings are set by a random singular value decomposition whose spectrum is linearly spaced with minimum-to-maximum ratio 0.8, normalized to unit mean squared singular value and scaled by the mean pretrained token norm; input and output embeddings are then frozen. We fine-tune Qwen2.5-0.5B with LoRA on the attention and MLP projections (q/k/v/o_proj and gate/up/down_proj), using rank 16, $\alpha = 32$, dropout 0.05, and bfloat16 training. For each of five seeds, we train on 1024 examples with AdamW (zero weight decay), learning rate 2×10^{-4}, batch size 4, gradient accumulation 8, maximum length 4096, and 2000 optimizer steps. Every 100 steps, we evaluate 256 examples per test depth. Training examples vary by seed, while the evaluation examples at each depth are shared across seeds. The plotted quantity is the mean token cross-entropy over the whole completion, including node identifiers, values, control tokens, the answer, and the end-of-sequence token, computed with teacher forcing. The reported curves average over five seeds and the shaded regions indicate one standard deviation.

Parallel updates and sequential traces. The theory writes all nodes of a given height in one CoT step, whereas the language model writes one node at a time. In the ground-truth height order, each node value depends only on values at smaller heights, which is the dependency structure of the height schedule, and producing each value requires routing analogous to that of the theory: locating the node's function and the values of its children among earlier tokens. Sequential generation differs in three respects. The model must also emit the identity of the next node, which the structural schedule fixes in the theory; causal attention can read the entire earlier prefix, including earlier records at the same height, rather than a shared embedding; and a tree with I internal nodes produces I node records and $4I + 3$ predicted tokens, including the answer and the end-of-sequence token, rather than one step per height. In addition, teacher forcing conditions every prediction on the true prefix, so the plotted loss does not include the compounding of errors in free-running generation. Thus Figure 4(b) measures conditional completion prediction on deeper serialized trees; it neither establishes the learned routing mechanism nor measures free-running reasoning accuracy.

C Proof of Theorem 1

We first state the auxiliary lemmas used in the proof.

Lemma 1. *Given $\{a_i\}_{i\in[N_1]} \subset \mathbb{R}^{m_1}$, $\{b_j\}_{j\in[N_2]} \subset \mathbb{R}^{m_2}$, where $a_1, \cdots, a_{N_1}$ are linearly independent, and $b_1, \cdots, b_{N_2}$ are linearly independent. Then for any fixed matrix $U \in \mathbb{R}^{N_1 \times N_2}$, there exists a matrix $W \in \mathbb{R}^{m_1 \times m_2}$ such that*

$$a_i^\top W b_j = U_{i,j},\ \forall i \in [N_1], j \in [N_2].$$

The proof of Lemma 1 is given in Appendix C.2.

Based on Lemma 1, we next prove that the induced bilinear form in the token subspace is unique.

Lemma 2. *Under the assumption of Lemma 1, suppose $W, \tilde{W} \in \mathbb{R}^{m_1 \times m_2}$ both satisfy*

$$a_i^\top W b_j = a_i^\top \tilde{W} b_j = U_{i,j}, \forall i \in [N_1], j \in [N_2].$$

Then for any $a \in \mathrm{span}\{a_i\}_{i\in[N_1]}$ and $b \in \mathrm{span}\{b_j\}_{j\in[N_2]}$, we have

$$a^\top W b = a^\top \tilde{W} b.$$

The proof of Lemma 2 is given in Appendix C.3.

C.1 Proof of Theorem 1

Order each node embedding as $\mathbf{E}_m = [x_m; r_m; f_m; v_m]$, where the block dimensions are d, d_r, d, d, respectively. Under the instantiation $\rho(y, l) = l \otimes y$, the relation subblock has dimension $d_p \cdot d$, where $d_p \geq N_c$ is the dimension of the positional encodings in $\mathcal{P}$, and $r_m = l_m \otimes y_m$. For the displayed block matrices, write $d_r = d_p \cdot d$; if the structural block also stores arity and height, pad the corresponding rows and columns of the score matrices by zeros. The immutable structural mask reads that metadata separately. We use the block order (x, r, f, v) for all matrices below, and write $0_{a\times b}$ for the zero matrix in $\mathbb{R}^{a\times b}$.

By Assumption 2, the public token matrix has full column rank. Applying Lemma 1 with the public tokens on both sides and target matrix $\beta_h I_{|\mathcal{B}|}$ gives, for each $h \in [N_c + 1]$, a matrix $B^h \in \mathbb{R}^{d\times d}$ satisfying

$$x^\top B^h x' = \begin{cases} \beta_h, & x = x', \\ 0, & x \neq x', \end{cases} \qquad \forall x, x' \in \mathcal{B}. \tag{19}$$

For each child head $h \in [N_c]$, define the value matrix

$$W_h^V = [0_{d\times d} \quad 0_{d\times d_r} \quad 0_{d\times d} \quad I_d], \tag{20}$$

so that $W_h^V \mathbf{E}_m = v_m$. Define the complete key-query matrix by

$$W_h^{KQ} = \begin{bmatrix} 0_{d\times d} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \\ q_h \otimes B^h & 0_{d_r\times d_r} & 0_{d_r\times d} & 0_{d_r\times d} \\ 0_{d\times d} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \end{bmatrix}. \tag{21}$$

Here $q_h \otimes B^h$ denotes the linear map $x \mapsto q_h \otimes (B^h x)$ from $\mathbb{R}^d$ to $\mathbb{R}^{d_r}$. Therefore, for source node m and target node n,

$$\alpha_{m,n}^{(h)} = \mathbf{E}_m^\top W_h^{KQ} \mathbf{E}_n = r_m^\top (q_h \otimes B^h x_n) = (l_m^\top q_h)(y_m^\top B^h x_n).$$

The hard slot mask restricts this head to $\mathcal{C}^{(h)} = \{m \neq o(\mathcal{G}) : l_m = q_h\}$; outside this pool its attention is zero. Inside the pool, $l_m^\top q_h = 1$ by Assumption 1. If $m = c_h(n)$, then $y_m = x_n$, so (19) gives $\alpha_{m,n}^{(h)} = \beta_h$. If $m \in \mathcal{C}^{(h)}$ and $m \neq c_h(n)$, the uniqueness of a child in each occupied slot implies $p(m) \neq n$. The injective identity assignment gives $y_m \neq x_n$, and hence $y_m^\top B^h x_n = 0$. This proves (7) on the eligible pool for every valid child slot. The raw zero score outside the pool is replaced by the hard mask and is never included in the softmax denominator.

For the function-retrieval head, define

$$W_{N_c+1}^V = [0_{d\times d} \quad 0_{d\times d_r} \quad I_d \quad 0_{d\times d}], \tag{22}$$

so that $W_{N_c+1}^V \mathbf{E}_m = f_m$, and define

$$W_{N_c+1}^{KQ} = \begin{bmatrix} B^{N_c+1} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \\ 0_{d_r\times d} & 0_{d_r\times d_r} & 0_{d_r\times d} & 0_{d_r\times d} \\ 0_{d\times d} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d_r} & 0_{d\times d} & 0_{d\times d} \end{bmatrix}. \tag{23}$$

Then

$$\alpha_{m,n}^{(N_c+1)} = \mathbf{E}_m^\top W_{N_c+1}^{KQ} \mathbf{E}_n = x_m^\top B^{N_c+1} x_n,$$

which equals β_{N_c+1} if $m = n$ and 0 otherwise by (19). This proves (8).

Finally, let $N_h = |\mathcal{C}^{(h)}|$ for child heads and $N_{N_c+1} = |\mathcal{V}|$ for the function head. For every active head and target node, the selected source has logit β_h and every other eligible source has logit zero. Thus its softmax mass is

$$\frac{\exp(\beta_h)}{\exp(\beta_h) + N_h - 1},$$

which tends to one as $\beta_h \to \infty$, while all other masses tend to zero. Because the child-head value matrices in (20) extract v_m and the function-head value matrix in (22) extracts f_m, the limiting attention outputs satisfy $\mathrm{attn}_h(O)_n \to v_{c_h(n)}$ for $h \in [N_c]$ and $\mathrm{attn}_{N_c+1}(O)_n \to f_n$. Therefore, in the large-margin limit, the constructed attention layer supplies exactly the ordered child values and function token required by the fixed calibrated semantic module for one bottom-up CoT update. Missing child slots are not queried, and calibration (M1) agrees with the intended node function on the resulting one-hot inputs.

C.2 Proof of Lemma 1

We let $A := (a_1, \cdots, a_{N_1}) \in \mathbb{R}^{m_1 \times N_1}, B := (b_1, \cdots, b_{N_2}) \in \mathbb{R}^{m_2 \times N_2}$. Then for any $W \in \mathbb{R}^{m_1 \times m_2}$, we have

$$A^\top W B = U \Leftrightarrow (B^\top \otimes A^\top) \cdot \mathrm{vec}(W) = \mathrm{vec}(U). \tag{24}$$

Since $a_1, \cdots, a_{N_1}$ are linearly independent, and $b_1, \cdots, b_{N_2}$ are linearly independent, we have $\mathrm{rank}(A) = N_1$ and $\mathrm{rank}(B) = N_2$, which implies

$$\mathrm{rank}(B^\top \otimes A^\top) = \mathrm{rank}(B^\top)\mathrm{rank}(A^\top) = N_1 N_2.$$

Therefore, the linear system

$$(B^\top \otimes A^\top) \cdot w = vec(U)$$

has a solution $w \in \mathbb{R}^{m_1 m_2}$. (24) implies reshaping w to a matrix $W \in \mathbb{R}^{m_1 \times m_2}$ gives the desired W.

C.3 Proof of Lemma 2

Define the subspaces

$$\mathcal{S}_A := \mathrm{span}\{a_1, \cdots, a_{N_1}\} \subset \mathbb{R}^{m_1},$$
$$\mathcal{S}_B := \mathrm{span}\{b_1, \cdots, b_{N_2}\} \subset \mathbb{R}^{m_2}.$$

Take any $a \in \mathcal{S}_A$ and $b \in \mathcal{S}_B$. By definition of the spans, there exist coefficient vectors $\alpha \in \mathbb{R}^{N_1}$ and $\beta \in \mathbb{R}^{N_2}$ such that

$$a = \sum_{i=1}^{N_1} \alpha_i a_i,\ b = \sum_{j=1}^{N_2} \beta_j b_j.$$

Let $A := (a_1, \cdots, a_{N_1}) \in \mathbb{R}^{m_1 \times N_1}$ and $B := (b_1, \cdots, b_{N_2}) \in \mathbb{R}^{m_2 \times N_2}$. Then we can write

$$a = A\alpha, b = B\beta.$$

First compute $a^\top W b$:

$$a^\top W b = (A\alpha)^\top W (B\beta) = \alpha^\top A^\top W B \beta.$$

By assumption in W, we have $A^\top W B = U$, so

$$a^\top W b = \alpha^\top U \beta.$$

Similarly, for $\tilde{W}$, we obtain $a^\top \tilde{W} b = \alpha^\top A^\top \tilde{W} B \beta$.

Since $\tilde{W}$ satisfies the same constraints, we also have $A^\top \tilde{W} B = U$, hence $a^\top W b = \alpha^\top U \beta = a^\top \tilde{W} b$ for all $a \in \mathcal{S}_A$ and $b \in \mathcal{S}_B$, which proves Lemma 2.

D PROOF OF PROPOSITION 1

Let $d(n)$ denote the structural height of node n, with leaves at height zero and $d(n) = 1 + \max_{m \in C(n)} d(m)$ at an internal node. Let the root $o(\mathcal{G})$ have height k. We first prove by induction on t that, in the large-margin limit of Theorem 1, after CoT step t every node n with $d(n) \le t$ has the correct value token in O^t.

For $t = 0$, the input embedding $O^0 = \mathbf{E}(\mathcal{G})$ contains the observed ground-truth values for all leaf nodes and zero value tokens for non-leaf nodes. Hence the claim holds for all nodes with height zero.

Assume the claim holds after step $t-1$. Consider a node n with $d(n) = t$, and let $H(n) \subseteq [N_c]$ be its valid child slots. For every $h \in H(n)$, the child $c_h(n)$ has height at most $t-1$, so the induction hypothesis places its correct value token in O^{t-1}. In the limiting construction of Theorem 1, child head h attends to $c_h(n)$ and the function-retrieval head attends to n itself. Consequently, the semantic module receives exactly the ordered list of existing child values and the correct function code. Inactive slots are not queried. By calibration (M1), these one-hot inputs produce the one-hot value token of the node function. The structural mask (M2) writes this value into node n at step t and freezes it thereafter; nodes of smaller height remain unchanged. Thus the induction claim holds at step t.

The argument also establishes convergence of the actual finite-margin rollout to this limiting computation. Every softmax on a finite eligible pool converges to its selected source as $\min_h \beta_h \to \infty$. The fixed semantic module is a finite multilinear polynomial in the soft function coefficients and operands, hence continuous. At each of the finitely many scheduled steps, attention averaging, semantic composition, and the fixed masked write therefore preserve convergence. Induction over these steps gives

$$\lim_{\min_h \beta_h \to \infty} \widehat{v}_{o(\mathcal{G})} = v_{o(\mathcal{G})}.$$

The structural schedule writes the root at step k, so this construction finishes a height-k tree in k CoT steps, with all nodes at the same height processed in parallel. This step count is a property of the specified schedule.

E CANONICAL MODEL AND NOTATION

The following table maps the notation used in these proofs to the manuscript.

Proof notation	Manuscript meaning
$B = [b_1, \dots, b_M]$	Public identity-token dictionary; $\sigma_{\min}(B) = \sigma_{\min}(B)$
B^h, $Z^h = B^\top B^h B$	Trainable combined attention matrix and its identity score table
z_n, $\Phi z_n = \widehat{v}_n$	Predicted categorical coefficients and the main-text predicted value block
f_n, Ψe_{f_n}	Node function and its encoded function-token block (main-text f_n)
y_n, $x_{p(n)}$	True symbol, with main-text value token $v_n = \Phi e_{y_n}$; parent identity token (main-text y_n), respectively
d_e	Identity embedding dimension, denoted d in the main text
G, $\mathcal{D}$	A tree and its population law, denoted $\mathcal{G}$ and $\mathcal{F}(\mathcal{G})$ in the main text
$\mathcal{I}$, I_*, $\overline{\mathcal{V}}$, K	Internal nodes and bounds on internal count, total count and height
$r = N_c + 1$, $\mathcal{A}$, ρ	Number of heads, structurally nontrivial heads and $\lvert\mathcal{A}\rvert$
$F(d)$, $\Lambda_h = -\partial_{d_h} F$	The original rollout loss restricted to score margins, and its drive
D, N_*, b_*	Unfolded attention-draw bound, pool-size bound and entry margin
W_h^{wit}, $W_{\min}$	Expected null-witness count and minimum across active heads

E.1 SYSTEM SETTING

We study tree-structured compositional reasoning as a bottom-up computation. Leaves carry input symbols. Each internal node applies a function from a finite dictionary to the ordered values of its children, and its result becomes an input to its parent. The learning target contains every internal value, including the root value, so the task includes the intermediate results of the computation.

The data distribution $\mathcal{D}$ consists of a semantic law over tree shapes, node functions, and leaf symbols, together with an assignment of identity tokens to nodes. Identity tokens identify nodes and their parents for attention routing; the function and value tokens encode the computation. The task class permits any finite alphabet, a finite function dictionary of bounded arity, bounded ordered rooted trees, and any semantic law. Additional distributional or geometric assumptions are stated with the results that require them.

An instance is an ordered rooted tree $G = (\mathcal{V}, \mathcal{E})$. Its root is o, its parent map is p, and its internal and leaf sets are $\mathcal{I}$ and $\mathcal{L}$. Each internal node n has arity $a_n \in \{1, \dots, N_c\}$ and children $c_1(n), \dots, c_{a_n}(n)$. Occupied slots are the prefix $1, \dots, a_n$. The root has no slot and is excluded from child pools. Leaves have height zero and $\mathrm{ht}(n) = 1 + \max_h \mathrm{ht}(c_h(n))$. Assume $|\mathcal{V}| \le \overline{\mathcal{V}}$, $|\mathcal{I}| \le I_*$, and $\mathrm{ht}(o) \le K$, with finite bounds $\overline{\mathcal{V}} \ge 2$, $I_* \ge 1$, and $K \ge 1$. Values belong to $[q]$. A finite dictionary $\mathcal{F}$ contains functions $f : [q]^{a_f} \to [q]$, with declared arities $a_f \le N_c$. The function at n has arity a_n. The true symbol is $y_n = f_n(y_{c_1(n)}, \dots, y_{c_{a_n}(n)})$. A semantic instance includes the shape, functions, and leaf symbols. Their joint law is arbitrary unless a result adds assumptions.

The identity dictionary is the full-column-rank matrix $B = [b_1, \dots, b_M] \in \mathbb{R}^{d_e \times M}$, with $M \ge \max\{2, \overline{\mathcal{V}}\}$. Write $C = \max_j \|b_j\|_2^2$. The execution expansion (Lemma 5), the structural barriers (Lemma 9 and the barrier inequalities of Corollary 2), and the proof of Theorem 4 in Appendix L never use this rank condition, so they hold for any dictionary B. An injective assignment $\iota : \mathcal{V} \to [M]$ gives $x_n = b_{\iota(n)}$ and the parent token $x_{p(n)}$. Uniform independent injection is assumed for the reduction theorems and, unless a result states a pair-probability substitute (Corollary 2 in Appendix J), for the zero-loss theorems. It is not required for stationarity, the smoothness bounds, or Lemma 10.

E.2 System model

The transformer uses multi-head attention for ordered child selection and function retrieval, followed by a fixed semantic module for local composition. Each node's input embedding has an identity block; a structural block containing its parent identity, child-slot indicator, arity, and height; a function-code block; and a value block. The structural and function blocks remain available throughout the computation, while the value block records observed leaves and subsequently generated intermediate results.

There is one child head for each possible slot and one function head. For a target node, the intended source of child head h is its h-th child, and the intended source of the function head is the target's own function code. At finite scores the reads are soft averages over the eligible pools defined below. A hard slot mask excludes sources in other slots. Each head has its own trainable combined bilinear score matrix B^h; all these matrices start at zero. The fixed module is specified by the soft-input calibration rule (M1) below.

Value tokens are Φe_ξ for symbols $\xi \in [q]$, with $\Phi^\top \Phi = I_q$. Function tokens are Ψe_f with $\Psi^\top \Psi = I_{|\mathcal{F}|}$. The blocks occupy separate coordinates. Null values and leaf function codes are zero. Write z_n for predicted categorical value coefficients. The corresponding main-text predicted value block is $\widehat{v}_n = \Phi z_n$.

We write O^i for the embedding after CoT step i, with O^0 the input embedding; the update is given in the CoT process below. In $O^0 = \mathbf{E}(G) = [X; R(G); F(G); U(G)]$, X has columns x_n. The block $R(G)$ concatenates parent identity tokens, slot indicators, and encodings of arities and heights; the root's parent and slot entries are zero. The function block has columns Ψe_{f_n} at internal nodes and zero at leaves. The value block $U(G)$ has columns Φe_{y_n} at leaves and zero at internal nodes. The target $O(G)$ keeps all three immutable blocks and replaces each internal value column by Φe_{y_n}.

The following four choices define the canonical model.

(M1) Calibrated fixed semantics. Write $\Delta^p_{\le 1} = \{x \in \mathbb{R}^p_{\ge 0} : \mathbf{1}^\top x \le 1\}$. At a target of arity a, independently draw a function code and a operands from $c \in \Delta^{|\mathcal{F}|}_{\le 1}$ and $u_h \in \Delta^q_{\le 1}$, completing each missing mass by a null symbol †. Apply the drawn truth table if its arity is a and every operand is non-null; otherwise return null, represented by $e_\dagger = 0$. Calibration means that the fixed module returns the non-null distribution of this independent

experiment for every such soft input. Thus

$$\mathsf{G}_a(c; u_1, \ldots, u_a)_v = \sum_{f: a_f = a} c_f \sum_{b \in [q]^a} \mathbf{1}\{f(b) = v\} \prod_{h=1}^{a} (u_h)_{b_h}. \tag{25}$$

Only existing operands are queried; null or incompatible codes and any null operand kill the output, including for constant functions.

(M2) Structural schedule. The immutable input includes arities, slots, and heights. At step i write exactly height-i nodes and freeze them afterwards. All reads use O^{i-1}. With nodes as columns, $M^i = \mathrm{diag}(\mathbf{1}\{\mathrm{ht}(n) = i\})$ multiplies on the right. The structural rule agrees with bottom-up evaluation at exact routing.

(M3) All-slot candidates with a hard slot mask. For $h \le N_c$, let $\mathcal{C}^{(h)} = \{m \ne o : \mathrm{slot}(m) = h\}$. For the function head $r = N_c + 1$, let $\mathcal{C}^{(r)} = \mathcal{V}$. The pools include unwritten nodes and null leaf function codes. Out-of-slot sources receive zero attention through a hard mask. The hard mask supplies column-shift invariance.

(M4) Orthonormal value and function coordinates. These make the value loss Euclidean in categorical coefficients. The value-loss arguments use this Euclidean metric. Identity tokens may be nonorthogonal.

For an existing child context, and for the function context, set

$$\begin{aligned} \alpha^{(h)}_{m,n} &= x_{p(m)}^\top B^h x_n && (h \le N_c, m \in \mathcal{C}^{(h)}), \\ \alpha^{(r)}_{m,n} &= x_m^\top B^r x_n && (m \in \mathcal{V}), \\ s^{(h)}_{m,n} &= \frac{\exp \alpha^{(h)}_{m,n}}{\sum_{j \in \mathcal{C}^{(h)}} \exp \alpha^{(h)}_{j,n}}. \end{aligned} \tag{26}$$

A queried child pool is nonempty because it contains the correct child. Empty pools are never queried. A singleton pool returns its sole value, independently of its score. A missing-child context is inactive. If its nonempty-pool softmax is recorded for bookkeeping, its output is unused. No incompatible function activates it.

E.3 The CoT process

Chain-of-Thought (CoT) is the iterative production and reuse of intermediate values in the embedding. The computation proceeds bottom-up, with all height-i nodes written in parallel at step i. A CoT step uses the following sequence:

1. Select the nodes of height i using the immutable structural schedule. Read only from the previous embedding O^{i-1}.
2. For each selected node, use one child head per existing slot to read a soft mixture of candidate value blocks, with its corresponding child as the intended source. Use the function head to read a soft mixture of function codes, with its own code as the intended source.
3. Apply the fixed semantic module to the retrieved operands and function coefficients, and write the resulting predicted value into that node's value block.
4. Freeze the newly written values. Preserve all other columns and all immutable blocks for the next step.

After K steps, the final embedding contains every predicted internal value, including the predicted root value. Write $\widehat{O}(G; \theta) = O^K$ for this final embedding, where $\theta = (B^1, \ldots, B^r)$ collects the trainable score matrices. The rollout is free running: later steps read the model's own earlier predictions from the embedding, without replacing them by ground-truth internal values. Equation (28) compares this final embedding with the ground-truth embedding, and Eq. (29) trains the separate score matrices by population gradient descent from zero.

At height $i = \mathrm{ht}(n)$ the operands and function coefficients are

$$u_{n,h} = \sum_{m \in \mathcal{C}^{(h)}} s^{(h)}_{m,n} z_m^{i-1}, \qquad c_n = \sum_{m \in \mathcal{V}} s^{(r)}_{m,n} e_{f_m}, \qquad z_n = \mathsf{G}_{a_n}(c_n; u_{n,1}, \ldots, u_{n,a_n}), \tag{27}$$

where $e_{f_m} = 0$ for a leaf. Here z_m^{i-1} denotes the value coefficients of node m in O^{i-1}. Initially leaves carry e_{y_m} and internal states are zero. The immutable blocks are copied exactly. Writing $\widetilde{Z}^i$ for the proposed value columns, the dimensionally correct update is $O^i = O^{i-1} + [0; 0; 0; \Phi\widetilde{Z}^i]M^i$. Every selected column was zero before that step.

For clarity, the objective and algorithm are restated here:

$$\mathcal{L}_{\text{train}}(\theta) = \frac{1}{2}\mathbb{E}_{G\sim\mathcal{D}}\|O(G) - \widehat{O}(G;\theta)\|_F^2, \tag{28}$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla_\theta\mathcal{L}_{\text{train}}(\theta^{(t)}), \qquad \theta^{(0)} = 0, \quad \theta = (B^1, \ldots, B^r). \tag{29}$$

These are combined bilinear matrices, not two separately zero-initialized query and key factors. The population and objective remain fixed.

E.4 Growth conventions

All implied constants below are absolute unless their dependence is stated. The semantic drive floors remain problem-dependent. For complete b-ary trees with $b \geq 2$,

$$I = \frac{b^K - 1}{b - 1} = \Theta(V/b), \qquad V = \frac{b^{K+1} - 1}{b - 1} = \Theta(b^K), \qquad D = (b + 1)I = \Theta(V). \tag{30}$$

For $I_\ell = (b^\ell - 1)/(b - 1)$, $I^2 \leq \Theta_2 = \sum_{\ell=1}^K b^{K-\ell} I_\ell^2 \leq [b/(b - 1)]I^2$, since $I_\ell/I \leq b^{\ell-K}$. Thus $\Theta_2 = \Theta(I^2)$. Complete-tree claims with nontrivial child heads assume $K \geq 2$.

F Supporting results for the exact reduction

This appendix supplies the module, orbit, execution, invariant-plane, and smoothness results used by the main reduction and convergence proofs.

F.1 Module, orbit, and execution formulas

Lemma 3 (Module and objective)**.** *Use the finite-alphabet canonical model (M1) through (M4), with bounded trees, any semantic law, and any injective identity assignment. The module maps subprobability inputs to a subprobability output. It is linear separately in each operand and in the function coefficients, and is exact on compatible one-hot inputs. Its surviving mass is*

$$\mathbf{1}^\top \mathsf{G}_a(c; u) = \left(\sum_{f:a_f=a} c_f\right) \prod_{h=1}^{a} \mathbf{1}^\top u_h. \tag{31}$$

It contracts the ℓ_1 norm in each argument when the others are fixed subprobability vectors. The rollout is well defined and

$$\mathcal{L}_{\text{train}}(\theta) = \frac{1}{2}\mathbb{E}\sum_{n\in\mathcal{I}}\|z_n - e_{y_n}\|_2^2. \tag{32}$$

All predictions here use the model's earlier predictions.

Proof. Every term in (25) is nonnegative. Summing over the output symbol counts each input tuple once and gives (31), which is at most one. Each monomial contains one entry of each operand and one function coefficient. At compatible one-hot inputs only the specified truth-table entry survives. In a fixed argument the map is a matrix with nonnegative entries and column sums at most one. For every signed vector v, the triangle inequality gives $\|Tv\|_1 \leq \|v\|_1$. Replacing arguments one at a time gives the sum of their ℓ_1 differences as a bound on the output difference. The tensor-product basis also proves uniqueness within the separately homogeneous multilinear class. Forming the monomials with product gates and summing their fixed truth-table coefficients realizes the map exactly. The number of operations is proportional to truth-table size times arity. No small representation of an arbitrary truth table is asserted.

Induct on the schedule. The initial leaf states and zero internal states are subprobability vectors. Attention produces convex combinations. The proved module property gives a subprobability write.

Only height-i columns are written at step i, so each internal node is written once and all earlier columns are preserved. Leaves and immutable blocks agree with the target. The remaining Frobenius discrepancy is the sum of squared value discrepancies. The identity $\Phi^\top \Phi = I_q$ then gives (32). □

Constants and growth rates. The argument-wise contraction constant is $L_{\arg} = 1 = \Theta(1)$, uniformly in q, arity, and tree size; the squared-loss normalization is exact. For truth-table size $T_a = |\{f : a_f = a\}| q^a$, the direct construction uses $O(aT_a)$ operations, since each entry contributes one monomial of arity a and its function coefficient. This is $O(bT_b)$ at each complete-tree node; no compressed truth-table assumption is made.

Lemma 4 (Distinct keys and exact orbit formula)**.** *In the canonical model on any fixed bounded ordered tree, the keys in each nonempty pool are distinct. In an active context exactly one key equals the target identity. Set $Z^h = B^\top B^h B$ and $R = I_M - \mathbf{1}\mathbf{1}^\top/M$. If $Z^h = d_h R$ for real d_h, then in a pool of size $N \geq 2$*

$$p(d; N) = \frac{e^d}{e^d + N - 1}, \qquad \gamma(d; N) = \frac{e^d - 1}{e^d + N - 1}, \qquad \gamma' = p(1 - \gamma). \tag{33}$$

The correct mass is p and every wrong mass is $(1 - \gamma)/N$. For $d \geq 0$ an active child read is the convex combination

$$u_{n,h} = \gamma_h z^{i-1}_{c_h(n)} + (1 - \gamma_h)\bar{z}_{h,i}, \qquad \bar{z}_{h,i} = \frac{1}{N_h} \sum_{m \in \mathcal{C}^{(h)}} z^{i-1}_m, \tag{34}$$

and $c_n = \gamma_r e_{f_n} + (1 - \gamma_r)\bar{c}$ with $\bar{c} = |\mathcal{V}|^{-1} \sum_m e_{f_m}$. A hypothetical missing-child context has uniform weights, independent of d_h. A singleton active pool always has mass one and zero derivative. For each shape the recursion in the fidelities is polynomial, with real coefficients, and with rational coefficients when the semantic law is rational.

Proof. Two slot-h children cannot have the same parent. Injectivity of identities therefore gives distinct parent keys in that pool. The correct child's parent is the target. Function keys are the node identities themselves. In $d_h R$, the correct logit is $d_h(1 - 1/M)$ and all wrong logits are $-d_h/M$. Subtracting the common offset leaves logits $d_h, 0, \ldots, 0$. Their softmax gives (33) and $p = \gamma + (1 - \gamma)/N$, proving (34). Differentiation of the quotient gives the derivative formula. If no diagonal key is present, all logits are equal, so the normalized law is uniform, not a law of total mass $1 - \gamma_h$. For a singleton the only softmax weight is one. One may set $p = 1, \gamma = 1 - e^{-d}$ algebraically; both endpoints in (34) coincide, so its derivative still vanishes. Successive substitutions into the finite multilinear recursion give a polynomial in the fidelities. Its averaging weights determine its coefficient field. For negative margins the algebraic formula remains valid, but the ideal-versus-uniform interpretation is not convex. □

Constants and growth rates. There are no additional tunable constants. The exact routing tail is $1 - p = (N - 1)/(e^d + N - 1) = \Theta(\min\{1, (N - 1)e^{-d}\})$ for $d \geq 0$: write it as $x/(1 + x)$ with $x = (N - 1)e^{-d}$ and bound it between $\min\{1, x\}/2$ and $\min\{1, x\}$. On complete trees the pools are $N_h = I = \Theta(V/b)$ and $N_r = V$.

Define $F(d)$ by the actual softmax for every $d \in \mathbb{R}^r$. For nonnegative margins, $F(d) = \mathbb{E}_\sigma F_\sigma(\gamma(d; \sigma))$. Neither identity assignments nor M enter this recursion. Shape-specific candidate counts do enter it.

Lemma 5 (Execution expansion and loss bounds)**.** *Use the bounded-tree, finite-alphabet canonical model with any semantic and injective identity laws, and any finite attention matrices. For a fixed semantic instance and identity assignment, there is a finite hard execution $Y_n \in [q] \cup \{\dagger\}$ with $e_\dagger = 0$ such that $z_n = \mathbb{E}[e_{Y_n} \mid G, \iota]$. Each execution uses at most*

$$S_K = \sum_{j=0}^{K-1} N_c^j, \qquad D = rS_K \tag{35}$$

semantic occurrences and attention draws, respectively, where the occurrence bound is S_K and the draw bound is D. With $\epsilon_n = 1 - z_{n,y_n}$ and $m_n = 1 - \mathbf{1}^\top z_n$,

$$\epsilon_n^2 \leq \|z_n - e_{y_n}\|_2^2 \leq 2\epsilon_n^2, \qquad \|z_n - e_{y_n}\|_1 = 2\epsilon_n - m_n \leq 2\epsilon_n. \tag{36}$$

On the margin orbit, for $d \geq 0$, let k_{nh} count head-h draws in the true subtree computation of output n, and let N_h^* bound its pool size over the support. Then

$$F(d) \leq \mathbb{E} \sum_n \left[\sum_h k_{nh}(N_h^* - 1)e^{-d_h} \right]^2 . \tag{37}$$

Also $F_\sigma(\gamma) \leq r^2 \Theta_2(\sigma) \Delta^2$, where $\Theta_2(\sigma) = \sum_n |\mathcal{I}(n)|^2$ and Δ is the largest $1 - \gamma_h$ among non-trivial active pools. Singleton draws may be omitted. In particular $\Theta_2 \leq \overline{\mathcal{V}}^3$.

Proof. At an internal occurrence, independently draw the function source and one source for each of the target's existing child slots, from their actual normalized softmax laws. A null or incompatible function causes †. A read of an unwritten internal state causes †. A leaf read returns its fixed symbol. Every written-state read unfolds into a fresh independent execution of that state, conditional on the entire instance. Return † if any operand is †; otherwise apply the compatible truth table. Heights decrease in every recursive call, so there are at most S_K occurrences and rS_K draws. Induction in height and summation of the independent draws reproduce (25). This includes mixed arities because incompatible codes are rejected and no missing slot is sampled. Independence belongs to the constructed execution copies, not to stored states or semantic variables.

The nontrue coordinates satisfy $\sum_{v \neq y_n} z_{n,v} = \epsilon_n - m_n \in [0, \epsilon_n]$. The true residual is $-\epsilon_n$. Squaring and using $\sum a_j^2 \leq (\sum a_j)^2$ for $a_j \geq 0$ proves (36). An all-correct execution visits the true subtree and returns the true output. Expose the first wrong choice on that path. A union bound gives $\epsilon_n \leq \sum_h k_{nh}(1 - p_h) \leq \sum_h k_{nh}(N_h^* - 1)e^{-d_h}$. The half-squared loss is at most ϵ_n^2, proving (37). Alternatively expose ideal-versus-uniform coins in (34). The event that every coin takes the ideal branch has probability $\prod_{m \in \mathcal{I}(n)} \gamma_r \prod_{h \leq a_m} \gamma_h$. It is contained in the all-correct event; the latter uses correct masses p_h, not fidelities. Its failure probability is at most $r|\mathcal{I}(n)|\Delta$. Squaring and summing gives the last bound. There are at most $\overline{\mathcal{V}}$ summands, each at most $\overline{\mathcal{V}}^2$. □

Constants and growth rates. The explicit draw bound is $D = rS_K$. If $N_c \geq 2$, $S_K = \Theta(N_c^{K-1})$ by the geometric-series bound; if $N_c = 1$, $S_K = K$. The loss-tail coefficient is $r^2\Theta_2 = O(r^2\overline{\mathcal{V}}^3)$, by the proved summand bound. On complete trees it is $\Theta((b+1)^2 I^2) = \Theta(V^2)$ by (30). The constants in the two semantic-error norm comparisons are absolute.

The loss is the squared error of the execution mean. It is not the expected squared error of a single execution. The bounds above concern semantic error probabilities. They do not lower-bound routing error.

F.2 Invariant plane and global smoothness

Lemma 6 (Invariant plane)**.** *Use the canonical model on a fixed complete ordered b-ary tree of height $K \geq 2$, with $b = N_c \geq 2$. Assume the joint semantic law is invariant under the following transformation for every permutation τ of the b slots: permute every path's slot labels by τ, transport all leaf values, and replace each transported gate by the correspondingly permuted truth table. Then $F(d_1, \ldots, d_b, d_f)$ is symmetric in its child margins on all of $\mathbb{R}^{b+1}$. In particular the invariance hypothesis holds for independent i.i.d. leaves and independent i.i.d. gates whose law is invariant under operand permutations. This includes every law on symmetric gates and the uniform law on all truth tables. With uniform independent identity injection,* (41)*, and separate-matrix population GD from zero, the reachable margins lie in*

$$\mathcal{P} = \{(d_c \mathbf{1}_b, d_f) : d_c, d_f \in \mathbb{R}\}. \tag{38}$$

Writing $\widetilde{F}(d_c, d_f) = F(d_c \mathbf{1}_b, d_f)$, the exact update is

$$d_c^+ = d_c - \frac{\alpha}{b} \partial_{d_c} \widetilde{F}, \qquad d_f^+ = d_f - \alpha \partial_{d_f} \widetilde{F}. \tag{39}$$

The plane child drive is $\Lambda_c = -(1/b)\partial_{d_c} \widetilde{F}$, the drive of each child head.

Proof. Represent each node by its path from the root. Mapping every slot on its path by τ is a bijection preserving heights and parent relations. It carries the slot-h pool onto the slot-$\tau(h)$ pool. Transporting a gate f to f' means $f'(u_{\tau(1)}, \dots, u_{\tau(b)}) = f(u_1, \dots, u_b)$ on symbols. Multilinearity gives this identity on soft inputs as well. By induction on height, the rollout on the transformed instance with permuted child margins is the transported original rollout. True outputs transform in the same way. Summing squared residuals and using invariance of the law proves symmetry of F. For the stated i.i.d. sufficient condition, reindexing does not change the product law of the leaves or gates, and replacing each gate by its operand permutation preserves its marginal law. Independence preserves the full law.

At a point with equal child margins, differentiating the symmetry identity for each child transposition makes all child partial derivatives equal. Theorem 5 therefore preserves equality of these margins by induction from zero. Finally the chain rule gives $\partial_{d_c}\widetilde{F} = \sum_{h=1}^{b} \partial_{d_h} F = b\partial_{d_1} F$ and $\partial_{d_f}\widetilde{F} = \partial_{d_r} F$. Substitution in (42) proves (39). No matrices were tied or projected. □

Constants and growth rates. The child and function scalar steps are exactly $\alpha/b = \Theta(\eta\lambda_\perp^2/(Mb))$ and $\alpha = \Theta(\eta\lambda_\perp^2/M)$. These rates follow directly from the chain-rule factor b and concern the complete-tree family of this lemma. The individual child-head step in the full coordinates remains α.

The corresponding symmetry argument for irregular shapes additionally uses invariance of the joint shape and slot law under the stated transformation.

Lemma 7 (Global smoothness)**.** *For the bounded-tree canonical model, arbitrary semantic and injective identity laws, and $C = \max_j \|b_j\|^2$, valid global Euclidean smoothness constants are*

$$L_\theta = 16 I_* C^2 D^2, \qquad L_d = 16 I_* D^2. \tag{40}$$

The first is for the product Frobenius norm of the raw matrices. The second is for all real margins in the actual softmax restriction. Also $0 \le F \le I_*$ and $|\partial_{d_h} \ell_G| \le 2I_* D$. Under the symmetry hypothesis of Lemma 6, each of the functions $(1/b)\partial_{d_c}\widetilde{F}$ and $\partial_{d_f}\widetilde{F}$ is L_d-Lipschitz in its own coordinate when the other plane coordinate is fixed.

Proof. Fix a unit raw-parameter direction U. Each score derivative has magnitude at most $\|b_j\|\|b_k\|\|U^h\|_F \le C$. For an execution weight w, its first log derivative has magnitude at most $2CD$. Its second log derivative is minus a sum of softmax variances of numbers in $[-C, C]$, so its magnitude is at most $C^2 D$. Differentiating the finite execution expansion gives $\|z_n'\|_1 \le 2CD$ and $\|z_n''\|_1 \le C^2(4D^2 + D)$. The residual norm is at most $\sqrt{2}$ by Lemma 5. The absolute second directional derivative of one half-squared loss is at most $4C^2D^2 + \sqrt{2}C^2(4D^2 + D) \le 16C^2D^2$ for $D \ge 1$. Summing and averaging bounds the symmetric Hessian operator norm. Integrating on line segments gives smoothness. For a unit margin direction use representative logits d_h for correct choices and zero for wrong choices. Their derivatives have magnitude at most one, giving L_d by the same calculation with $C = 1$. For a single margin coordinate the log derivative of each draw has magnitude at most one, whence $\|z_n'\|_1 \le D$ and the loss partial is bounded by $\sqrt{2} I_* D \le 2 I_* D$. Each output loss is at most one, proving the loss bound.

Let $w = (\mathbf{1}_b, 0)$ and let $\mathcal{H} = \nabla_d^2 F$ on the plane. The derivative of $(1/b)\partial_{d_c}\widetilde{F}$ in d_c equals $w^\top \mathcal{H} w/b$. Its magnitude is at most $L_d\|w\|^2/b = L_d$. The function-coordinate second derivative has magnitude at most L_d. Integration in the respective scalar coordinates proves the final assertion. □

Constants and growth rates. The displayed valid bounds satisfy $L_\theta = \Theta(I_* C^2 D^2)$ and $L_d = \Theta(I_* D^2)$; the per-head gradient bound is $O(I_* D)$ and the loss bound is $O(I_*)$. On complete trees these become $\Theta(C^2V^3/b)$, $\Theta(V^3/b)$, $O(V^2/b)$, and $O(V/b)$, respectively. They are obtained by summing the per-output directional estimates in the proof and substituting (30).

F.3 Calibrated realization

Proposition 2 (Unique affine calibration and shared product realization)**.** *For every finite alphabet and arity-declared finite dictionary,* (25) *is the unique calibrated map. It is also the unique separately affine map on $\Delta_{\le 1}^{|\mathcal{F}|} \times (\Delta_{\le 1}^{q})^a$ taking value $e_{f(b)}$ at compatible non-null product vertices*

and zero at every other product vertex, including null and incompatible vertices. For $q \geq 2$, the operand monomials can be computed with at most $P(a) \leq 2q^a$ binary product gates and product depth $\lceil \log_2 a \rceil$. A fixed arithmetic network then realizes the module using the linear map $S_{b,v} = \sum_{f:a_f=a} c_f \mathbf{1}\{f(b) = v\}$, q^{a+1} further binary products, and a linear sum over b. Put $Q = \max\{q, 2\}$. For all $a \leq N_c$, the total size, counting product gates and nonzero coefficients in the linear maps, is at most $\mathcal{S}_{\mathrm{mod}} := (3N_c + |\mathcal{F}|)Q^{N_c+1}$ and the depth is at most $\lceil \log_2 N_c \rceil + 3$, counting unbounded-fan-in linear layers. The case $q = 1$ embeds in two symbols. Every soft output is the conditional mean of the fresh-copy unfolded hard execution of Lemma 5 (Appendix F); Eq. (28) *is the squared error of that mean.*

Constants and growth rates. The explicit size bound is $\mathcal{S}_{\mathrm{mod}} = \Theta((N_c + |\mathcal{F}|)Q^{N_c+1})$ and the depth bound is $\Theta(1 + \log N_c)$, with absolute implied constants. For a complete b-ary tree the shared-module bounds are $\Theta((b + |\mathcal{F}|)Q^{b+1})$ and $\Theta(1 + \log b)$, independently of height K. These are bounds on the displayed realization, not lower bounds on the minimum circuit size. They follow by summing the product recurrence and the linear coefficients counted in the proof.

Proof. The independent experiment in (M1) gives (25) by multiplication of its independent draw probabilities, and defines the output uniquely. Each subprobability simplex has vertices $0, e_1, \ldots, e_p$ and the affine barycentric representation $x = (1 - \mathbf{1}^\top x)0 + \sum_j x_j e_j$. Apply separate affinity successively in every input block. All vertex terms with a null block or incompatible code have prescribed value zero; the remaining terms are precisely (25). Conversely that polynomial has the prescribed values.

Split a operands into groups of sizes $\lceil a/2 \rceil$ and $\lfloor a/2 \rfloor$, compute all group monomials recursively, then multiply every pair of group monomials. The count obeys

$$P(1) = 0, \qquad P(a) = q^a + P(\lceil a/2 \rceil) + P(\lfloor a/2 \rfloor).$$

For $a = 2, 3$ the bound is immediate. If $a \geq 4$, both group sizes are at least two, so $2q^{\lceil a/2 \rceil} + 2q^{\lfloor a/2 \rfloor} \leq q^a$: after division by q^a, the left side is at most $2/4 + 2/4 = 1$. Induction gives $P(a) \leq 2q^a$. The depth recursion gives $\lceil \log_2 a \rceil$. The linear code map can be computed alongside these monomials or in one additional layer; one product layer and one output-sum layer give the stated depth. Over all arities, operand products cost at most $2N_cQ^{N_c}$, further products at most $N_cQ^{N_c+1}$, and the output sums use at most $N_cQ^{N_c+1}$ nonzero coefficients. The code maps use at most $|\mathcal{F}|Q^{N_c+1}$ nonzero coefficients. Since $Q \geq 2$, their sum is at most $(3N_c+|\mathcal{F}|)Q^{N_c+1} = \mathcal{S}_{\mathrm{mod}}$. No compression of an arbitrary truth table is asserted.

Condition on the complete semantic instance and identity assignment. Whenever an execution reads a written prediction, unfold a fresh independent execution for that read. Heights decrease, so this terminates. Induction using calibrated composition and independent attention draws shows that its output indicator has conditional mean z_n. Stored predictions and semantic variables need not be independent. Substitution into (32) (Appendix F) proves the squared-mean statement; reusing a single sampled stored value would define a different execution. □

Remark 2 (Smooth approximation)**.** *The finite product realization admits uniform smooth-network approximation on each compact input set. For example, choose a smooth activation $\sigma = \tanh$ and a point a with $\sigma''(a) \neq 0$. Taylor's theorem shows that*

$$S_h(t) = \frac{\sigma(a + ht) + \sigma(a - ht) - 2\sigma(a)}{h^2\sigma''(a)} \longrightarrow t^2$$

uniformly with its first derivative on bounded intervals as $h \to 0$. The polarization identity $uv = ((u + v)^2 - (u - v)^2)/4$ therefore supplies a smooth-network approximation to each product gate, uniformly with first derivatives. Composing the finitely many gates preserves this convergence on compact input sets. The theorems assume exact calibration; this approximation observation makes no claim about training an approximate module.

G EXACT REDUCTION OF GD AND ARBITRARY DICTIONARY GEOMETRY

Margin and execution notation. Write $Z^h = B^\top B^h B$ and $R = I_M - \mathbf{1}\mathbf{1}^\top/M$. For a pool of size $N \geq 2$, the correct mass and fidelity on the score tuple $Z^h = d_h R$ are denoted by

$$p(d;N) = \frac{e^d}{e^d + N - 1}, \qquad \gamma(d;N) = \frac{e^d - 1}{e^d + N - 1}.$$

Here N_h^σ is the head-h pool size on shape σ; singleton reads are already exact. Let $F(d)$ denote the actual-softmax population loss on this tuple for every real d. For a fixed shape, $F_\sigma(\gamma)$ denotes the conditional loss of the recursion whose reads interpolate the correct-source and uniform-pool reads with fidelities γ_h, as specified in Lemma 4 (Appendix F). For nonnegative margins the notation is $F(d) = \mathbb{E}_\sigma F_\sigma(\gamma(d;\sigma))$. Write $\mathcal{I}(n)$ for the internal nodes in the true subtree of n, and let k_{nh} count head-h draws in its true subtree computation. The execution notation of Lemma 5 (Appendix F) is

$$S_K = \sum_{j=0}^{K-1} N_c^j, \qquad D = rS_K, \qquad \Theta_2(\sigma) = \sum_{n\in\mathcal{I}} |\mathcal{I}(n)|^2, \qquad \Theta_2 = \mathbb{E}_\sigma \Theta_2(\sigma).$$

The smoothness constants used below are $L_\theta = 16 I_* C^2 D^2$ and $L_d = 16 I_* D^2$, as in Lemma 7 (Appendix F). The first result identifies the exact population updates under an equicorrelated identity Gram; the second gives the stationarity conclusion for a general full-rank Gram.

Theorem 5 (A: exact population reduction)**.** *Use the canonical model on arbitrary bounded ordered trees, with any joint finite-alphabet semantic law. Let identities be assigned by a uniform injection independently of that law and the shape. Assume the full-rank dictionary satisfies*

$$G_B = B^\top B = \lambda_\perp I_M + \beta \mathbf{1}\mathbf{1}^\top, \qquad \lambda_\perp > 0, \quad \lambda_\perp + M\beta > 0. \tag{41}$$

For population GD (29) *on separate matrices from zero and every finite $\eta > 0$, the identity score tables satisfy $Z_t^h = d_{h,t} R$ and*

$$d_{h,0} = 0, \qquad d_{h,t+1} = d_{h,t} + \alpha \Lambda_h(d_t), \quad \Lambda_h = -\partial_{d_h} F, \quad \alpha = \frac{\eta \lambda_\perp^2}{M-1} \geq \frac{\eta \sigma_{\min}(B)^4}{M-1}. \tag{42}$$

No drive condition or equality between different heads is assumed. At every orbit point the aggregated score gradient is

$$\Gamma_h^Z := \nabla_{Z^h} \mathcal{L}_{\text{train}} = \frac{\partial_{d_h} F}{M-1} R, \qquad \nabla_{B^h} \mathcal{L}_{\text{train}} = B \Gamma_h^Z B^\top. \tag{43}$$

Proof. The population loss depends on the parameters only through the score tables. A permutation of dictionary identities simultaneously conjugates all tables and bijects the uniform injections. Hence the population loss is invariant under this conjugation. Adding $\mathbf{1}c^\top$ to a table adds one constant to all eligible scores in each target column. Every masked softmax and the whole rollout are unchanged. At a tuple $(d_h R)_h$, conjugation fixes the tuple. Differentiating its invariance shows that each partial gradient has a common diagonal entry and a common off-diagonal entry. It therefore lies in $\mathrm{span}\{I_M, \mathbf{1}\mathbf{1}^\top\}$. Column-shift invariance gives zero column sums, leaving a multiple of R. The derivative along $d_h R$ is its inner product with R. Since $\|R\|_F^2 = M - 1$, this determines (43). The matrix chain rule follows from $Z_{jk}^h = b_j^\top B^h b_k$.

The raw update induces $Z^{h,+} = Z^h - \eta G_B \Gamma_h^Z G_B$. The identities $R\mathbf{1} = 0$ and (41) give $G_B R G_B = \lambda_\perp^2 R$. Starting at zero, induction proves (42) for every real margin reached. The eigenvalues of G_B are $\lambda_\perp$ and $\lambda_\perp + M\beta$. Thus $\sigma_{\min}(B)^2 \leq \lambda_\perp$, proving the last inequality. For normalized nonnegative equicorrelation, $G_B = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$, $0 \leq \rho < 1$, equality holds. □

Constants and growth rates. The exact effective step obeys $\alpha = \Theta(\eta \lambda_\perp^2 / M)$ for $M \geq 2$, with $\alpha \geq \eta \sigma_{\min}(B)^4/(M-1)$. The aggregated-gradient multiplier $(M-1)^{-1} = \Theta(M^{-1})$ follows from $\|R\|_F^2 = M - 1$. These formulas also apply to complete trees, with $M \geq V$; if $M = V$, the effective step is $\Theta(\eta \lambda_\perp^2 / V)$.

Proposition 3 (G: general raw Gram gives stationarity)**.** *Use the bounded-tree finite-alphabet canonical model, any semantic law, any injective identity-assignment law, and any full-column-rank identity dictionary B. Run population GD* (29) *with $0 < \eta < 2/L_\theta$, where* (40) *(Appendix F) defines L_θ. For the population score-table gradients $\Gamma^Z_{h,t}$,*

$$\min_{0\le t<T} \sum_h \|\Gamma^Z_{h,t}\|_F^2 \le \frac{\mathcal{L}_{\mathrm{train}}(0)}{\eta(1-\eta L_\theta/2)\sigma_{\min}(B)^4 T}. \tag{44}$$

The raw parameter gradient tends to zero. No zero-loss conclusion follows.

Proof. The matrix chain rule after population aggregation gives $\nabla_{B^h}\mathcal{L}_{\mathrm{train}} = B\Gamma^Z_h B^\top$. For a thin singular-value decomposition $B = U\Sigma V^\top$, unitary invariance of Frobenius norm and entrywise multiplication by singular values give $\|B\Gamma B^\top\|_F^2 \ge \sigma_{\min}(B)^4\|\Gamma\|_F^2$. Smoothness gives $\mathcal{L}_{\mathrm{train}}(\theta^+) \le \mathcal{L}_{\mathrm{train}}(\theta) - \eta(1-\eta L_\theta/2)\sum_h \|B\Gamma^Z_h B^\top\|_F^2$. Sum for $t < T$ and use nonnegativity of the final loss. The minimum is at most the mean, proving (44). The same telescoping bound over every finite horizon bounds the infinite sum of squared raw gradient norms. Its nonnegative summands tend to zero. □

Constants and growth rates. Writing $\chi_\theta = 1 - \eta L_\theta/2 > 0$, the score-gradient bound is $O(I_*/(\eta\chi_\theta\sigma_{\min}(B)^4 T))$, hence $O(V/(b\eta\chi_\theta\sigma_{\min}(B)^4 T))$ on complete trees. This uses $\mathcal{L}_{\mathrm{train}}(0) \le I_*$ in (44). The admissible step endpoint is $\Theta((I_* C^2 D^2)^{-1})$, or $\Theta(b/(C^2V^3))$ on that family. No uniform lower bound on χ_θ is implicit when the step approaches its endpoint.

H Convergence and certified semantic families

H.1 Any-step convergence under a corrective drive

For the complete-tree, slot-symmetric specialization, use the plane $\mathcal{P} = \{(d_c\mathbf{1}_b, d_f) : d_c, d_f \in \mathbb{R}\}$ and write $\widetilde{F}(d_c, d_f) = F(d_c\mathbf{1}_b, d_f)$, $\Lambda_c = -(1/b)\partial_{d_c}\widetilde{F}$, and $\Lambda_f = -\partial_{d_f}\widetilde{F}$, as in Lemma 6 (Appendix F). The following conditions distinguish global corrective drive from a drive floor needed only before basin entry.

A head is *structurally nontrivial* on a shape if it is queried and its pool has at least two members. Let $\mathcal{A}$ be the union of these heads over supported shapes. Other heads never change a read and can be omitted from loss and routing bounds. This definition does not use semantic influence. A constant gate can depend on an operand through survival.

The solved case without nontrivial heads. The set $\mathcal{A}$ is empty exactly when every supported tree is a single leaf. Indeed, an internal node n queries the function head, whose pool $\mathcal{C}^{(r)} = \mathcal{V}$ contains n and its first child, so that head is structurally nontrivial. This includes unary chains and trees whose child pools are singletons. Conversely, a single leaf queries no head. When $\mathcal{A} = \varnothing$, no supported instance has an internal value column, so the objective (28) vanishes identically, its gradient is zero, and population GD stays at zero. In every convergence statement and growth-rate formula below, this solved case has iteration budget zero. Entry-time maxima, floor, witness and certificate minima, and coefficients divided by $\rho = |\mathcal{A}|$ are defined and used only when $\mathcal{A} \ne \varnothing$. The class of tasks is unchanged.

Definition 2 (Uniform corrective drive)**.** *For a canonical population, condition (D) with constant $c > 0$ means that for every supported shape, every $h \in \mathcal{A}$, and all relevant fidelities,*

$$-\partial_{\gamma_h} F_\sigma(\gamma) \ge c(1-\gamma_h). \tag{D}$$

The box version quantifies over $[0,1]^{|\mathcal{A}|}$. For a complete b-ary tree with the slot-symmetric law of Lemma 6 (Appendix F), the plane version quantifies only over $\gamma_1 = \cdots = \gamma_b \in [0,1]$ and $\gamma_f \in [0,1]$. The derivative is that of an individual head before restriction.

This is a sufficient functional condition. It is not an assertion that all standard populations satisfy it. For rational finite laws it is a polynomial positivity problem. Nonnegative coefficients provide a sufficient test; a finite grid does not. The coefficient certificates in Appendix K use their stated polynomial bases.

Theorem 6 (B: global convergence under (D), any learning rate)**.** *Assume the canonical model, uniform independent identity injection, the equicorrelated full-rank Gram* (41)*, and zero-initialized population GD on separate matrices with any finite $\eta > 0$. Suppose each supported shape has $N_h^\sigma \geq 2$ for each $h \in \mathcal{A}$. Assume either the box version of (D), or a fixed complete b-ary shape, the slot symmetry in Lemma 6 (Appendix F), and the plane version of (D). Then every structurally nontrivial margin is nondecreasing and diverges, the loss is nonincreasing, and*

$$e^{2d_{h,t}} \geq 1 + \frac{2\alpha c t}{\overline{\mathcal{V}}}, \tag{45}$$

$$F(d_t) \leq \frac{r^2 \Theta_2 \overline{\mathcal{V}}^2}{1 + 2\alpha c t/\overline{\mathcal{V}}}, \qquad \Theta_2 = \mathbb{E}_\sigma \sum_n |\mathcal{I}(n)|^2. \tag{46}$$

A sufficient integer budget for loss at most $\varepsilon > 0$ is

$$T^* = \left\lceil \frac{r^2 \Theta_2 \overline{\mathcal{V}}^3 (M-1)}{2\eta \lambda_\perp^2 c\varepsilon} \right\rceil \leq \left\lceil \frac{r^2 \overline{\mathcal{V}}^6 (M-1)}{2\eta \sigma_{\min}(B)^4 c\varepsilon} \right\rceil. \tag{47}$$

Singleton contexts are already exact. Heads never used in a nontrivial context remain at zero.

Proof. Theorem 5 gives the exact margin update. In the plane case Lemma 6 (Appendix F) preserves the plane. For $u = e^{d_h} \geq 1$ and a supported count $N \in [2, \overline{\mathcal{V}}]$, (D) and the chain rule give a contribution at least

$$c\, p_h (1 - \gamma_h)^2 = c \frac{N^2 u}{(u + N - 1)^3} \geq \frac{c}{N u^2} \geq \frac{c}{\overline{\mathcal{V}} u^2}.$$

The first inequality uses $u + N - 1 \leq Nu$. Averaging shapes preserves this lower bound. Thus $d_h^+ - d_h \geq \alpha c e^{-2d_h}/\overline{\mathcal{V}} > 0$. Starting at zero proves nonnegativity and applicability of the hypothesis at every step. In the box case the loss is nonincreasing along each positive coordinate segment. In the plane case it is nonincreasing along each positive common-child or function-coordinate segment because the common-child derivative is the sum of equal nonpositive partials. Connecting successive iterates by those segments proves loss monotonicity without a step-size restriction.

Set $u_t = e^{d_{h,t}}$. The inequality $e^v \geq 1 + v$ implies $u_{t+1}^2 \geq u_t^2 + 2\alpha c/\overline{\mathcal{V}}$. Induction proves (45). For every supported count, $1 - \gamma_h = N/(e^{d_h} + N - 1) \leq \overline{\mathcal{V}} e^{-d_h}$. Use the last bound of Lemma 5 (Appendix F) to obtain (46). The stated ceiling is a sufficient, conservative solution of that inequality. The margin lower bound diverges with t, so each correct mass tends to one. An unused or singleton-only head has zero loss derivative and stays at zero. □

Constants and growth rates. For $t > 0$, the certified loss bound is $O(r^2 \Theta_2 \overline{\mathcal{V}}^3/(\alpha c t))$, and the displayed budget is $O(1 + r^2 \Theta_2 \overline{\mathcal{V}}^3/(\alpha c \varepsilon))$. These follow by dropping the one in (46) and using $\lceil x \rceil \leq 1 + x$. Since $\Theta_2 \leq \overline{\mathcal{V}}^3$, the general budget is also $O(1 + r^2 \overline{\mathcal{V}}^6 M/(\eta \sigma_{\min}(B)^4 c\varepsilon))$. For complete trees, (30) gives the sharper $O(1 + V^5/(\alpha c\varepsilon)) = O(1 + MV^5/(\eta\lambda_\perp^2 c\varepsilon))$. The corresponding loss rate is $O(MV^5/(\eta\lambda_\perp^2 ct))$. The dependence on c is explicit and is not bounded by these tree counts.

H.2 Universal large-margin inequality and conditional basin entry

The supporting execution-gradient estimate is Lemma 8 (Appendix I). Use $N_* \geq 2$ as a bound on all candidate counts and $Q(d) = D(N_* - 1) e^{-\min_{h \in \mathcal{A}} d_h}$ for its error coefficient, with $Q = 0$ if $\mathcal{A}$ is empty.

The explicit threshold $b_* = \log(2D(N_* - 1))$ gives $-\sum_h \partial_{d_h} F \geq F$ whenever every structurally nontrivial margin is at least b_*. Heads outside $\mathcal{A}$ make only singleton draws or none, so they affect neither Q nor this derivative. This is a common-direction inequality. It does not alone prove that GD enters or stays in the large-margin region. The next theorem supplies entry and invariance under one positivity condition on a pre-basin region.

Theorem 7 (C: convergence from zero under a pre-basin drive floor)**.** *Use the zero-vector canonical model on bounded ordered trees with any finite semantic law, uniform independent identity injection into a full-rank dictionary satisfying* (41)*, and zero-initialized population GD on separate matrices. Let $\mathcal{A}$ be the structurally nontrivial heads, let $N_* \geq 2$ bound all candidate counts, and put $L_d = 16 I_* D^2$ and $b_* = \log(2D(N_* - 1))$. For $h \in \mathcal{A}$ define the drive floor*

$$g_h = \inf\{\Lambda_h(d) : \ d \in [0,\infty)^r, \ d_h \leq b_*\}. \tag{48}$$

Assume one of the following two conditions.

- *(P)* $g_h > 0$ *for every $h \in \mathcal{A}$. In this case set $\rho = |\mathcal{A}|$ and let the index j below range over $\mathcal{A}$.*
- *($P_{\mathcal{P}}$) The tree is a fixed complete b-ary tree with $b \geq 2$ and $K \geq 2$, the semantic law satisfies the slot symmetry of Lemma 6 (Appendix F), and for $j \in \{c, f\}$*

  $$g_j = \inf\{\Lambda_j(d_c \mathbf{1}_b, d_f) : \ d_c, d_f \geq 0, \ d_j \leq b_*\} > 0,$$

  with Λ_c and Λ_f the plane drives of Lemma 6 (Appendix F). In this case set $\rho = b + 1$; here $N_* = V$, $D = (b+1)I$, and $I_* = I$.

If $0 < \alpha = \eta \lambda_\perp^2/(M-1) \leq 1/L_d$, GD stays in the nonnegative orthant, and under ($P_{\mathcal{P}}$) on the nonnegative plane. Each relevant margin crosses b_* permanently by

$$T_1 = \max_j \left\lceil \frac{b_*}{\alpha g_j} \right\rceil. \tag{49}$$

The region where all relevant margins are at least b_* is forward invariant. For every integer $s \geq 0$, if $F_{T_1} > 0$,

$$F_{T_1+s} \leq \left[F_{T_1}^{-1} + \frac{\alpha(1 - \alpha L_d/2)}{\rho} s \right]^{-1}. \tag{50}$$

A sufficient budget for error $\varepsilon > 0$ is $T_1 + \lceil \rho/(\alpha(1 - \alpha L_d/2)\varepsilon) \rceil$. Zero loss, if attained, remains zero. If $\mathcal{A}$ is empty, all reads are exact, so the loss and the required budget are zero.

Proof. Theorem 5 gives the exact margin update $d_h^+ = d_h + \alpha \Lambda_h(d)$. Under ($P_{\mathcal{P}}$), Lemma 6 (Appendix F) keeps the iterates on the plane and gives (39) (Appendix F). Write $q_j = -\Lambda_j$ for the relevant coordinates; on the plane these are $q_c = (1/b)\partial_{d_c} \widetilde{F}$ and $q_f = \partial_{d_f} \widetilde{F}$. Each q_j is L_d-Lipschitz in its own coordinate when the other coordinates are fixed. For a single margin coordinate this holds because L_d bounds the Hessian operator norm in Lemma 7 (Appendix F); on the plane it is the last assertion of that lemma. Heads outside $\mathcal{A}$ have zero partial derivative and keep zero margin.

Nonnegativity. Let the current point be nonnegative and replace only its coordinate j by zero. On the plane this replaces all child margins together and stays on the plane. The new point lies in the set defining g_j, so its value of q_j is at most $-g_j$. The own-coordinate Lipschitz bound gives $q_j \leq -g_j + L_d d_j$ at the current point. Hence $d_j^+ \geq (1 - \alpha L_d) d_j + \alpha g_j \geq \alpha g_j > 0$. Induction from zero proves the invariance.

Permanent crossing. If $d_j \geq b_*$, replace coordinate j by b_* instead. The new point again lies in the set defining g_j, and the same bound gives $d_j^+ - b_* \geq (1 - \alpha L_d)(d_j - b_*) + \alpha g_j > 0$.

Entry time. While $d_j < b_*$, the current nonnegative point lies in the set defining g_j. Each step therefore adds at least αg_j to d_j. Starting at zero, $d_j \geq b_*$ holds after at most $\lceil b_*/(\alpha g_j) \rceil$ steps and persists afterwards. The maximum over j proves (49) and the forward invariance.

Rate. In the entered region every relevant margin is at least b_*. Singleton draws have no error probability, so Lemma 8 (Appendix I) applies with $Q \leq D(N_* - 1)e^{-b_*} = 1/2$ and gives $-\sum_{h \in \mathcal{A}} \partial_{d_h} F \geq F$. The margin gradient has at most ρ nonzero coordinates, so Cauchy Schwarz gives $\|\nabla_d F\|^2 \geq F^2/\rho$. On the plane the gradient has b equal child coordinates q_c and one function coordinate q_f, and $(bq_c + q_f)^2 \leq (b+1)(bq_c^2 + q_f^2)$. The update is the full margin GD step, so the L_d descent inequality gives $F_{t+1} \leq F_t - \alpha(1 - \alpha L_d/2) F_t^2/\rho$. For positive consecutive losses, $(1/F_{t+1}) - (1/F_t) = (F_t - F_{t+1})/(F_t F_{t+1}) \geq \alpha(1 - \alpha L_d/2)/\rho$, because $F_{t+1} \leq F_t$. Summation proves (50) and the stated budget. A zero squared-residual loss has zero gradient and is fixed. □

Constants and growth rates. Put $g_{\min} = \min_j g_j$. The stated entry bound is $\Theta(1 + b_*/(\alpha g_{\min}))$, since a ceiling lies between $\max\{1, x\}$ and $1 + x$. Thus it is $O(1 + (1 + \log D + \log N_*)/(\alpha g_{\min}))$. The post-entry loss is $O(\rho/(\alpha s))$ for $s > 0$ because $1 - \alpha L_d/2 \in [1/2, 1)$, and its accuracy budget is $O(1+\rho/(\alpha\varepsilon))$. For complete trees these become $\Theta(1+\log V/(\alpha g_{\min}))$ and $O(b/(\alpha s))$, respectively, with allowed step $\alpha \le \Theta(b/V^3)$. The latter notation describes the explicit endpoint $1/L_d$, not a new step hypothesis.

The two phases are a proof device. The population, loss, module, and optimizer never change. The floor constrains a head only while its margin is below b_*, and its proof uses the stated smoothness bound above that threshold.

Corollary 1 (Corrective drive implies the drive floor)**.** *Keep the setting of Theorem 7. Suppose each supported shape has $N_h^\sigma \ge 2$ for every $h \in \mathcal{A}$. If the box version of (D) holds with constant c, then (P) holds with*

$$g_h \ge \frac{ce^{-2b_*}}{\overline{\mathcal{V}}} = \frac{c}{4\overline{\mathcal{V}}D^2(N_* - 1)^2}.$$

For a fixed complete tree with the slot symmetry of Lemma 6, the plane version of (D) similarly implies $(P_{\mathcal{P}})$.

Proof. At any point defining the floor, put $u = e^{d_h} \in [1, e^{b_*}]$. By (D) and the derivative in (33), each shape's contribution to Λ_h is at least

$$c\, p_h(1 - \gamma_h)^2 = \frac{cN^2u}{(u + N - 1)^3} \ge \frac{c}{Nu^2} \ge \frac{ce^{-2b_*}}{\overline{\mathcal{V}}}.$$

Averaging shapes preserves this lower bound. The individual child derivative on the invariant plane satisfies the same identity. □

The floor lower bound has order $c/(\overline{\mathcal{V}}D^2N_*^2)$, which is $\Theta(c/V^5)$ on complete trees.

Remark 3 (Checking the drive floor)**.** *Put $z_h = e^{-d_h} \in (0, 1]$. Every attention weight is a rational function of z whose denominator is a product of factors $1 + (N - 1)z_h \ge 1$, and at $z_h = 0$ the weights are those of exact routing. Hence F and each drive $\Lambda_h = z_h\partial_{z_h}F$ extend continuously to the closed cube $[0, 1]^r$. The set defining g_h is mapped onto a dense subset of the compact set $\{z_h \in [e^{-b_*}, 1],\ z_{h'} \in [0, 1]$ for $h' \ne h\}$. Therefore g_h is the minimum of the continuous extension over that compact set, and $g_h > 0$ if and only if $\Lambda_h > 0$ at every point of it, including points at which the other heads route exactly. On the plane the compact set is a rectangle in (z_c, z_f). For a rational finite law, (P) is thus strict positivity of finitely many rational functions on a compact box, which exact interval or Bernstein subdivision can certify. Condition (P) constrains the total drive, not an upper bound on its adverse part. It is implied by (D), and it tolerates adverse higher-level contributions whenever the total drive stays positive.*

I Execution-gradient inequality

I.1 The execution-gradient inequality

Lemma 8 (Execution-gradient inequality)**.** *Use the bounded-tree canonical model with any semantic law, and let $N_* \ge 2$ bound all candidate counts. For all real margin vectors,*

$$-\sum_{h=1}^{r} \partial_{d_h}F(d) \ge 2(1 - Q(d))F(d), \qquad Q(d) = D(N_* - 1)e^{-\min_{h\in\mathcal{A}} d_h}. \tag{51}$$

Set $Q = 0$ if $\mathcal{A}$ is empty. Also $|\sum_h \partial_{d_h}F| \le 2I_*D$. Off the orbit, the same inequality holds with left side $-\sum_h\langle\Gamma_h^Z, R\rangle$ and $Q = D(N_* - 1)e^{-b_*}$, if every active correct-versus-wrong score gap is at least b_*. Singleton draws have no error probability.

Proof. Increase all margins at unit speed. An execution τ with $J(\tau)$ wrong choices has $(\log w_\tau)' = \sum_{e\in\tau}(1-p_e) - J(\tau)$. Each error event contributing to ϵ_n or to a nontrue coordinate $z_{n,a}$ has at least one wrong choice by Lemma 5. Also $\sum_e(1 - p_e) \le Q(d)$. Consequently the negative derivative

of each of these nonnegative error probabilities is at least $(1 - Q)$ times that probability. Since $F = \frac{1}{2}\mathbb{E}\sum_n(\epsilon_n^2 + \sum_{a\neq y_n} z_{n,a}^2)$, multiply each probability inequality by that probability and sum to obtain (51). Each draw has log derivative of magnitude at most one in this direction. The finite execution expansion gives $\|z_n'\|_1 \le D$, so the absolute loss derivative is at most $\sqrt{2}I_*D \le 2I_*D$. Off the orbit, adding R raises each diagonal-versus-off-diagonal gap at unit speed and changes no missing-child context. The same log-derivative identity holds with the actual softmax correct mass. The gap assumption bounds each wrong probability by $(N_* - 1)e^{-b_*}$, proving the asserted extension. □

Constants and growth rates. The error coefficient is $Q = \Theta(DN_*e^{-\min_{h\in\mathcal{A}} d_h})$ for $N_* \ge 2$ and $\mathcal{A} \neq \varnothing$; the total-drive bound is $O(I_*D)$. On complete trees they are $\Theta(V^2e^{-\min_{h\in\mathcal{A}} d_h})$ and $O(V^2/b)$. The choice $b_* = \log(2D(N_* - 1))$ is $\Theta(1 + \log D + \log N_*)$, or $\Theta(\log V)$ on that family. These assertions use $N_*/2 \le N_* - 1 < N_*$ and (30).

J Structural barriers and general token dictionaries

Structural witness probabilities. For $h \in \mathcal{A}$, a *null witness* in a shape is an active target n and a wrong candidate m whose input is identically zero when n is written: for a child head, m occupies slot h and $\mathrm{ht}(m) \ge \mathrm{ht}(n)$; for the function head, m is a leaf. Put $\pi_h = \mathbb{P}(\text{the shape has such a head-}h\text{ witness})$ and $\pi_{\min} = \min_{h\in\mathcal{A}} \pi_h$. This is a structural probability, independent of semantic influences. The function-head probability is $\pi_r = \mathbb{P}(|\mathcal{I}| > 0)$; on complete trees of height at least two every active head has probability one. The counted and bounded-tree versions appear in Lemma 9 (Appendix J) and Corollary 2 (Appendix J).

For these witnesses, W_h^{wit} is their expected count and $W_{\min} = \min_{h\in\mathcal{A}} W_h^{\mathrm{wit}}$. For any score tuple put

$$g_{jk}^h = Z_{kk}^h - Z_{jk}^h, \qquad m(Z) = \min_{h\in\mathcal{A},\, j\neq k} g_{jk}^h, \qquad \mathcal{L}(Z) = \mathcal{L}_{\mathrm{train}}(Z).$$

The minimum is always restricted to structurally nontrivial heads.

J.1 Structural witnesses, irregular trees, and score estimates

Lemma 9 (W: counted structural witnesses control every score gap)**.** *Use the zero-vector canonical module, hard all-slot masks, the structural height schedule, orthonormal value coordinates, and the loss on all internal outputs. The shape is a fixed complete b-ary tree with $b = N_c \ge 1$, height $K \ge 2$, and an arbitrary finite-alphabet semantic law. Identity injection into $[M]$ is uniform and independent of semantics, where $M \ge V$. Set*

$$S_a = \sum_{j=0}^{a-1} b^j, \quad I = S_K, \quad V = I + b^K, \quad r = b + 1,$$

$$N_{\mathrm{child}} = I, \qquad N_{\mathrm{fun}} = V,$$

$$W_{\mathrm{child}}^{\mathrm{wit}} = \sum_{a=1}^{K-1} b^a S_a > 0, \qquad W_{\mathrm{fun}}^{\mathrm{wit}} = Ib^K.$$

For each child head let $(N_h, W_h^{\mathrm{wit}}) = (N_{\mathrm{child}}, W_{\mathrm{child}}^{\mathrm{wit}})$; for the function head let $(N_r, W_r^{\mathrm{wit}}) = (N_{\mathrm{fun}}, W_{\mathrm{fun}}^{\mathrm{wit}})$. For every finite score table,

$$\mathcal{L}(Z) \ge \min_h \frac{W_h^{\mathrm{wit}}}{2M(M-1)(e^{m(Z)} + N_h - 1)^2}. \tag{52}$$

Also, if every gap is at least a, then for every head and ordered pair $j \neq k$,

$$\mathcal{L}(Z) \ge \frac{W_h^{\mathrm{wit}} e^{-2g_{jk}^h}}{2M(M-1)(1 + (N_h - 1)e^{-a})^2}. \tag{53}$$

The coercivity comes from the null witnesses supplied by (M1) through (M3), $M \ge V$, uniform injection, and supervision of every internal output. No Gram symmetry, correlation-sign or semantic-influence condition, or from-zero premise is needed.

Constants and growth rates. For $b \geq 2$, the exact witness counts satisfy

$$W^{\mathrm{wit}}_{\mathrm{child}} = \Theta(I^2/b) = \Theta(V^2/b^3), \qquad W^{\mathrm{wit}}_{\mathrm{fun}} = \Theta(IV) = \Theta(V^2/b).$$

To derive the first order, the last term of the witness sum is $b^{K-1}S_{K-1}$; each preceding term is at most b^{-2} times its successor, since $S_{a+1} = 1 + bS_a$. Thus the sum is between that last term and its multiple $(1-b^{-2})^{-1}$. Uniformly for $b \geq 2$, $S_{K-1} = \Theta(b^{K-2})$ and $I = \Theta(b^{K-1})$. The second order follows from $b^K = \Theta(V)$. Consequently the head-specific coefficient in (53) is $\Theta(W^{\mathrm{wit}}_h/[M^2(1 + (N_h - 1)e^{-a})^2])$. For the unary family the formulas instead give $I = K$, $V = K + 1$, $W^{\mathrm{wit}}_{\mathrm{child}} = K(K-1)/2 = \Theta(K^2)$ and $W^{\mathrm{wit}}_{\mathrm{fun}} = K$; the exact displayed bounds apply to it as well.

Proof. At an internal target n, a slot-h candidate u whose height is at least that of n is still zero when n is written. If its attention probability is s, the hth operand has mass at most $1 - s$. The canonical survival identity makes the output mass at most $1 - s$, regardless of the truth table or other inputs. Its true-coordinate deficit is therefore at least s, and its contribution to the loss is at least $s^2/2$. A leaf in a function pool supplies the identical argument because its function code is zero.

For a child head, choose a target n of height $i \in \{1, \ldots, K-1\}$ and a parent v of height at least $i + 1$; its slot-h child is such an unwritten candidate. There are b^{K-i} targets and S_{K-i} possible parents, giving $W^{\mathrm{wit}}_{\mathrm{child}}$ ordered witnesses. In every witness $v \neq n$. For the function head, every (internal target, leaf) pair is a witness, giving $W^{\mathrm{wit}}_{\mathrm{fun}}$ witnesses. Assigning the wrong key identity j and target identity k to one fixed witness has probability $1/(M(M-1))$. For fixed (j, k), these events are mutually exclusive across witnesses: the injection identifies at most one wrong-key node and one target node. Thus their nonnegative lower bounds can be summed before averaging.

Given the witness assignment, the selected wrong probability is

$$s = \frac{e^{-g^h_{jk}}}{1 + \sum_{\ell \neq k:\, \ell \text{ in pool}} e^{-g^h_{\ell k}}}.$$

If all gaps are at least a, its denominator is at most $1 + (N_h - 1)e^{-a}$. Squaring, multiplying by the witness probability, and summing proves (53). To prove (52), take a head and pair attaining $m(Z)$ and use $a = m(Z)$. The resulting wrong probability is at least $1/(e^{m(Z)} + N_h - 1)$. The minimum over heads supplies a bound independent of the minimizing head. □

Corollary 2 (Bounded irregular trees and pair exposure)**.** *Use the bounded-tree canonical model (M1) through (M4), any joint law of shapes and finite-alphabet semantics, and $M \geq \max_\sigma |\mathcal{V}_\sigma|$. Conditional on the semantic instance, assign identities uniformly over injections. Let $\mathcal{A} \neq \varnothing$ contain the structurally nontrivial heads, $\rho = |\mathcal{A}|$, and $N_* \geq 2$ bound their pool sizes. Let W^{wit}_h be the expected number of head-h null witnesses (n, m) in the sense of the structural witness definition above, and π_h the probability defined there. Assume $W^{\mathrm{wit}}_h > 0$ for every $h \in \mathcal{A}$, equivalently $\pi_h > 0$. For arbitrary finite score tables, with m taken over $\mathcal{A}$,*

$$\mathcal{L}(Z) \geq \frac{\min_{h \in \mathcal{A}} W^{\mathrm{wit}}_h}{2M(M-1)(e^{m(Z)} + N_* - 1)^2}, \tag{54}$$

$$\mathcal{L}(Z) \geq \frac{\pi_h}{2M(M-1)(e^{m_{hk}} + N_* - 1)^2}, \qquad m_{hk} = \min_{j \neq k} g^h_{jk}. \tag{55}$$

In particular $\pi_r = \mathbb{P}(|\mathcal{I}| > 0)$; it equals one only when an internal node exists almost surely. Complete unary trees are included by $I = S_K = \sum_{a=0}^{K-1} b^a$ at $b = 1$. The conclusions of Theorem 8 also hold for this class with $D = rS_K$, I_* as in Section E, $L_\theta = 16 I_* C^2 D^2$, $0 < \eta < 2/L_\theta$, ρ in place of r in κ_B, every gap, m and selected mass taken over heads in $\mathcal{A}$, $W^{\mathrm{wit}}_{\min} = \min_{h \in \mathcal{A}} W^{\mathrm{wit}}_h$, $N_* - 1$ in place of $V - 1$ in the divergence bound, and

$$b_* = \log(2D(N_* - 1)), \qquad \phi_* = \frac{W^{\mathrm{wit}}_{\min}}{2M(M-1)(2D+1)^2(N_* - 1)^2}.$$

Heads outside $\mathcal{A}$ have zero gradient, so their tables stay fixed; the theorem makes no claim about their gaps.

For a nonuniform injection law, a sufficient substitute is a constant $p_{\mathrm{pair}} > 0$ such that, conditional on every supported semantic instance and each of its null witnesses (n, m) (for the π_h bound, only

the witness selected by the fixed structural rule), every ordered pair $j \neq k$ satisfies $\mathbb{P}(\iota(\kappa) = j, \iota(n) = k \mid \text{instance}) \geq p_{\text{pair}}$, where $\kappa = p(m)$ for a child head and $\kappa = m$ for the function head. Then replace $1/[M(M-1)]$ everywhere in these bounds and in ϕ_* by p_{pair}; the factors W_h^{wit} and π_h remain unconditional shape expectations and probabilities.

Constants and growth rates. Writing $W_{\min} = \min_{h \in \mathcal{A}} W_h^{\text{wit}}$, the certified basin threshold has order

$$\phi_* = \Theta\left(\frac{W_{\min}}{M^2 D^2 N_*^2}\right), \qquad \kappa_B = \Theta\left(\frac{\eta(1 - \eta L_\theta/2)\sigma_{\min}(B)^4}{\rho M}\right).$$

The column-barrier coefficient is $\Theta(\pi_h/[M^2(e^{m_{hk}} + N_* - 1)^2])$. These comparisons use only $D \geq 1$, $N_* \geq 2$, $M \geq 2$ and the displayed formulas. For a nonuniform injection replace M^{-2} by p_{pair} in the barrier and threshold orders; the rate coefficient is unchanged. On complete trees with $b \geq 2, K \geq 2$, Lemma 9 gives $W_{\min} = \Theta(V^2/b^3)$, whence $\phi_* = \Theta((M^2 b^3 V^2)^{-1})$ and $\kappa_B = \Theta(\eta(1 - \eta L_\theta/2)\sigma_{\min}(B)^4/(bM))$. The post-entry loss is $O((\kappa_B s)^{-1})$ for $s > 0$.

Proof. Condition on the semantic instance and repeat the disjoint-witness sum in Lemma 9, replacing the pool size by N_*. Average the witness count to get the first inequality. For the second, choose one witness by a fixed structural rule whenever one exists, and choose j^* maximizing $Z_h[jk]$ over $j \neq k$. On the event assigning j^* to its key and k to its target, every other wrong candidate scores at most $Z_h[j^*k]$. Thus the chosen null weight is at least $1/(e^{m_{hk}} + N_* - 1)$, and the node loss is at least half its square. Its unconditional event probability is $\pi_h/[M(M-1)]$. Every finite nonempty tree with an internal node has a leaf, proving the assertion for π_r. Bounded witness counts give $W_h^{\text{wit}} > 0 \iff \pi_h > 0$. The conditional pair lower bound gives the same two averaging arguments with p_{pair} and the stated factors, without ambiguity about conditioning.

At $m = b_*$ the right side of (54) equals ϕ_* and is strictly decreasing in m. Thus $\mathcal{L} \leq \phi_*$ implies $m \geq b_*$ over $\mathcal{A}$. Lemma 11, with $D = rS_K$ and the internal-node bound I_* from Section E, gives $L_\theta = 16 I_* C^2 D^2$. Its raw descent bound preserves this sublevel for $0 < \eta < 2/L_\theta$. The common score direction supported on $\mathcal{A}$ has squared norm $\rho(M-1)$, and Lemma 10 gives $-\sum_{h \in \mathcal{A}} \langle \Gamma_h^Z, R \rangle \geq \mathcal{L}$ there. Cauchy Schwarz and the singular-value bound for B therefore give

$$\sum_{h \in \mathcal{A}} \|B \Gamma_h^Z B^\top\|_F^2 \geq \frac{\sigma_{\min}(B)^4 \mathcal{L}^2}{\rho(M-1)}.$$

Combining this with raw descent gives $\mathcal{L}_{t+1} \leq \mathcal{L}_t - \kappa_B \mathcal{L}_t^2$ with the stated replacement of r by ρ. Reciprocal summation, as in the proof of Theorem 8 below, gives the rate bound (59) with these constants. Solving (54) for m and inserting that rate proves the divergence bound with $W_{\min}^{\text{wit}}$ and $N_* - 1$. Its logarithm has positive argument because $\mathcal{L}_{t_0} \leq \phi_*$ makes that argument at least $2D(N_* - 1) > 0$. Every queried selected mass of an active head is $(1 + \sum_{j \neq k,\, j \text{ in pool}} e^{-g_{jk}^h})^{-1}$ and tends to one. The same argument uses p_{pair} in the nonuniform case. Heads outside $\mathcal{A}$ are unqueried or have only singleton queried pools; their tables do not change any read, so their gradients vanish and their tables stay fixed. A structurally nontrivial head with $W_h^{\text{wit}} = 0$ needs another argument and is excluded by the hypothesis. □

Lemma 10 (O: off-orbit execution-gradient inequality)**.** *Use the bounded-tree canonical model, any semantic law and any injective identity law. Let $N_* \geq 2$ bound candidate counts. If every gap of every structurally nontrivial head is at least $\underline{g} \in \mathbb{R}$, then*

$$-\sum_{h \in \mathcal{A}} \langle \Gamma_h^Z, R \rangle \geq 2(1 - Q)\mathcal{L}(Z), \qquad Q = D(N_* - 1)e^{-\underline{g}}.$$

Thus the left side is at least $\mathcal{L}(Z)$ when $\underline{g} \geq b_*$.

Constants and growth rates. The exact tail constant is $Q = D(N_* - 1)e^{-\underline{g}} = \Theta(DN_* e^{-\underline{g}})$, so $b_* = \log(2D(N_* - 1)) = \Theta(\log(2DN_*))$ is sufficient for the unit loss coefficient. On complete trees $Q = \Theta(V^2 e^{-\underline{g}})$ and $b_* = \Theta(\log V)$. Both follow by substitution in the execution count in the proof.

Proof. Along $Z_h + sR$ every active gap grows at unit speed. For correct mass p_e, correct and wrong log weights have derivatives $1 - p_e$ and $-p_e$. An unfolded execution with J_{wrong} wrong draws has log-weight derivative $\sum_e (1 - p_e) - J_{\text{wrong}} \le Q - J_{\text{wrong}}$. An all-correct execution returns the true symbol, so every event defining a true-coordinate deficit or a wrong-coordinate probability has at least one wrong draw. Its probability p therefore satisfies $p' \le -(1 - Q)p$. The loss is half the sum of squares of those nonnegative probabilities; differentiating gives the assertion. Inactive and singleton-only heads contribute zero. □

Lemma 11 (Score smoothness and raw descent)**.** *For the bounded-tree canonical model with any semantic and injective identity laws, let $D = rS_K$ and I_* bound the number of internal nodes. Valid global bounds are*

$$L_Z = 16 I_* D^2, \qquad \|\nabla \mathcal{L}(Z)\| \le J = 2 I_* D, \qquad \|\nabla \mathcal{L}(Z)\| \le 2D\sqrt{2 I_* \mathcal{L}(Z)}.$$

There are also relative bounds, with $H_{\text{rel}} = 16D^2 + 2D$:

$$\|\nabla \mathcal{L}(Z)\| \le 4D\mathcal{L}(Z), \qquad \|\nabla^2 \mathcal{L}(Z)\|_{\text{op}} \le H_{\text{rel}} \mathcal{L}(Z). \tag{56}$$

For any fixed full-rank B, put $G_B = B^\top B$, $C = \max_j \|b_j\|^2$, and $L_\theta = 16 I_* C^2 D^2$. One raw GD step obeys, for $0 < \eta < 2/L_\theta$,

$$\mathcal{L}(Z^+) \le \mathcal{L}(Z) - \eta(1 - \eta L_\theta/2) \sum_h \|B \Gamma_h^Z B^\top\|_F^2. \tag{57}$$

Alternatively the score-space proof gives the same bound with L_θ replaced by $L_Z \lambda_{\max}(G_B)^2$ and $\|B\Gamma_h^Z B^\top\|_F^2 = \langle \Gamma_h^Z, G_B \Gamma_h^Z G_B \rangle$.

Constants and growth rates. The displayed formulas give $L_Z = \Theta(I_* D^2)$, $L_\theta = \Theta(C^2 I_* D^2)$, $J = \Theta(I_* D)$, square-root-loss coefficient $2D\sqrt{2I_*} = \Theta(D\sqrt{I_*})$, relative gradient coefficient $4D = \Theta(D)$, and $H_{\text{rel}} = \Theta(D^2)$. For complete trees these are, respectively, $\Theta(V^3/b)$, $\Theta(C^2 V^3/b)$, $\Theta(V^2/b)$, $\Theta(V^{3/2}/\sqrt{b})$, $\Theta(V)$ and $\Theta(V^2)$. The raw step ceiling is $\Theta((C^2 I_* D^2)^{-1})$, or $\Theta(b/(C^2 V^3))$ on that family. The proof derives each coefficient from at most D log-weight derivatives and at most I_* loss terms.

Proof. For a unit score-table direction, every logit derivative has magnitude at most one. An execution contains at most D attention draws. Its first log-weight derivative is bounded by $2D$, and its second log-weight derivative is minus a sum of softmax variances, bounded in absolute value by D. The execution expansion therefore gives $\|z_n'\|_1 \le 2D$ and $\|z_n''\|_1 \le 4D^2 + D$. The residual has Euclidean norm at most $\sqrt{2}$, so the absolute second directional derivative of one half-squared output loss is at most

$$4D^2 + \sqrt{2}(4D^2 + D) \le 16D^2.$$

Summing and averaging bounds the symmetric Hessian operator norm and proves L_Z-smoothness. Pairing residuals and output tangents by Cauchy Schwarz over outputs and the population also gives $\|\nabla \mathcal{L}\| \le \sqrt{2\mathcal{L}}\, 2D\sqrt{I_*}$. Since every residual coordinate has magnitude at most one, the first directional derivative of each half-squared output loss has magnitude at most $2D$, proving the global gradient bound J. For a unit raw-parameter direction the corresponding bound on each logit derivative is C, yielding L_θ by the same calculation. The matrix chain rule gives $\nabla_{B^h} \mathcal{L} = B\Gamma_h^Z B^\top$; the usual smoothness descent inequality proves (57). For a proof entirely in scores, the positive operator $\mathcal{P}_{G_B} : U \mapsto (G_B U_h G_B)_h$ has norm $\lambda_{\max}(G_B)^2$, and $\|\mathcal{P}_{G_B}\Gamma\|^2 \le \lambda_{\max}(G_B)^2 \langle \Gamma, \mathcal{P}_{G_B}\Gamma \rangle$.

For the relative bounds, express the loss as half the sum of squares of nonnegative error-event probabilities: the true-coordinate deficit and each wrong-coordinate probability. Every such event is a union of hard executions, so the same log-weight bounds give, in any unit direction, $|p'| \le 2Dp$ and $|p''| \le (4D^2 + D)p$. Consequently the loss first derivative is at most $4D\mathcal{L}$ in absolute value, and its second derivative is at most $(16D^2 + 2D)\mathcal{L}$. Supremizing over unit directions proves (56). □

Theorem 8 (A loss basin for arbitrary full-rank tokens)**.** *Assume exactly the complete-tree canonical model and population hypotheses of Lemma 9 (Appendix J), with any full-column-rank identity dictionary B. Put $D = rI$, $I_* = I$, $N_* = V$, and*

$$b_* = \log(2D(V-1)), \qquad \phi_* = \min_h \frac{W_h^{\text{wit}}}{2M(M-1)(2D(V-1) + N_h - 1)^2}. \tag{58}$$

Run ordinary population GD on separate raw matrices with $0 < \eta < 2/L_\theta$, where $L_\theta = 16IC^2D^2$ and $C = \max_j \|b_j\|^2$. Suppose that at some iterate t_0, $\mathcal{L}(Z_{t_0}) \le \phi_*$. Then every later iterate has every gap at least b_*, remains in the same loss sublevel, and converges to zero loss. More precisely, writing $\sigma_{\min} = \sigma_{\min}(B)$ and

$$\kappa_B = \frac{\eta(1-\eta L_\theta/2)\sigma_{\min}^4}{r(M-1)},$$

one has, when $\mathcal{L}(Z_{t_0}) > 0$,

$$\mathcal{L}(Z_{t_0+s}) \le [\mathcal{L}(Z_{t_0})^{-1} + \kappa_B s]^{-1} \qquad (s \in \mathbb{Z}_{\ge 0}). \tag{59}$$

No score-gap premise in addition to the loss premise is required. This is a basin theorem, and does not assert that arbitrary full-rank dictionaries enter the basin from zero.

Moreover, writing $W_{\min}^{\mathrm{wit}} = \min_h W_h^{\mathrm{wit}}$, every gap diverges to infinity, although individual gaps need not be monotone; every selected mass tends to one. For every $s \ge 0$ (the argument of the logarithm is positive in the basin),

$$m(Z_{t_0+s}) \ge \log\left(\sqrt{\frac{W_{\min}^{\mathrm{wit}}}{2M(M-1)}(\mathcal{L}(Z_{t_0})^{-1} + \kappa_B s)} - (V-1)\right).$$

Constants and growth rates. Let $W_{\min} = \min_h W_h^{\mathrm{wit}}$. Since $N_h \le V$ and $D \ge 1$, the denominators defining ϕ_* are uniformly comparable to $M^2D^2V^2$. Hence

$$\phi_* = \Theta\left(\frac{W_{\min}}{M^2D^2V^2}\right), \qquad \kappa_B = \Theta\left(\frac{\eta(1-\eta L_\theta/2)\sigma_{\min}(B)^4}{rM}\right).$$

For $b \ge 2$, the witness estimates of Lemma 9 (Appendix J) give $\phi_* = \Theta((M^2b^3V^2)^{-1})$, $b_* = \Theta(\log V)$, $L_\theta = \Theta(C^2V^3/b)$, and $\kappa_B = \Theta(\eta(1-\eta L_\theta/2)\sigma_{\min}(B)^4/(bM))$. The post-entry loss is $O((\kappa_B s)^{-1})$ for $s > 0$; $O(1 + (\kappa_B\varepsilon)^{-1})$ further steps suffice for loss at most $\varepsilon > 0$. No uniform lower bound on $1 - \eta L_\theta/2$ is implicit in these orders. For unary trees the exact witness formulas instead give $\phi_* = \Theta((M^2K^3)^{-1})$.

Proof. The right side of (52) (Appendix J) is strictly decreasing in m. Its value at $m = b_*$ is ϕ_*. Thus $\mathcal{L}(Z) \le \phi_*$ forces $m(Z) \ge b_*$: if $m < b_*$, each head's lower bound is strictly greater than its value at b_* and the minimum is strictly greater than ϕ_*. Raw descent preserves the loss sublevel, so it preserves all these gap bounds without any restriction on individual gap increments.

For completeness, the common score direction is $A = (R, \dots, R)$. Increasing Z by sA raises every correct-versus-wrong gap at unit speed. An execution with J_{wrong} wrong choices has log-weight derivative $\sum_e (1 - p_e) - J_{\mathrm{wrong}}$. Each error event has at least one wrong choice. In the stated region, $\sum_e (1 - p_e) \le D(V-1)e^{-b_*} = 1/2$. Every nonnegative error probability therefore decreases at a rate at least half its own value in direction A. Squaring the true-coordinate deficit and the wrong-coordinate probabilities gives

$$-\sum_h \langle \Gamma_h^Z, R\rangle \ge \mathcal{L}(Z).$$

Since $\|A\|^2 = r(M-1)$, Cauchy Schwarz yields $\sum_h \|\Gamma_h^Z\|_F^2 \ge \mathcal{L}(Z)^2/[r(M-1)]$. The singular-value decomposition of B gives

$$\sum_h \|B\Gamma_h^Z B^\top\|_F^2 \ge \sigma_{\min}^4 \sum_h \|\Gamma_h^Z\|_F^2.$$

Insert these inequalities into (57) (Appendix J). The result is $\mathcal{L}_{t+1} \le \mathcal{L}_t - \kappa_B \mathcal{L}_t^2$. For positive consecutive losses this implies $\mathcal{L}_{t+1}^{-1} - \mathcal{L}_t^{-1} \ge \kappa_B$; summing proves (59). A zero squared-residual loss has zero gradient and stays zero. In fact the structural lower bound excludes zero loss at any finite score table in this setting. Finally, (52) (Appendix J) implies the weaker common-denominator bound $\mathcal{L}(Z) \ge W_{\min}^{\mathrm{wit}}/[2M(M-1)(e^{m(Z)} + V - 1)^2]$. Solving for $m(Z)$ and inserting (59) proves the displayed divergence bound. To check positivity, choose h_0 with $W_{h_0}^{\mathrm{wit}} = W_{\min}^{\mathrm{wit}}$. Since $\mathcal{L}(Z_{t_0}) \le \phi_*$, the logarithm's argument is at least $2D(V-1) + N_{h_0} - V > 0$. In every queried pool, $p = (1 + \sum_{j\ne k,\, j \text{ in pool}} e^{-g_{jk}^h})^{-1}$ therefore tends to one, uniformly over the finite set of pools and identity assignments. □

K CERTIFIED HEIGHT-TWO ILLUSTRATION

K.1 FINITE COEFFICIENT CERTIFICATES AND A LABELLED BOOLEAN ILLUSTRATION

Theorem 9 (Algebraic certificate)**.** *Use the bounded-tree canonical model, uniform independent identity injection, the full-rank Gram* (41)*, and separate-matrix population GD from zero. Put* $x_h = e^{d_h} - 1$*. Suppose the fixed loss is* $F = A(x)/Q_0(x)$*, where* Q_0 *is a positive product of powers of factors* $N + x_h$ *on* $x \geq 0$*. For every structurally nontrivial head suppose* $\delta_h > 0$ *and*

$$(1 + x_h)^3 (A\partial_h Q_0 - Q_0 \partial_h A) - \delta_h Q_0^2 \quad \textit{has nonnegative coefficients.} \tag{60}$$

Any omitted head must have no nontrivial context on the supported tasks. Then every finite $\eta > 0$ *gives monotone margins and loss, and*

$$e^{-2d_{h,t}} \leq (1 + 2\alpha\delta_h t)^{-1}, \tag{61}$$

$$F_t \leq \min \left\{ F_0, \mathbb{E} \sum_n \left[\sum_h \frac{k_{nh}(N_h^* - 1)}{\sqrt{1 + 2\alpha\delta_h t}} \right]^2 \right\}. \tag{62}$$

Such a rational representation exists for every finite semantic population; arbitrary real population weights give real coefficients.

Constants and growth rates. Set $\delta_{\min} = \min_{h \in \mathcal{A}} \delta_h$ and $A_{\text{cert}} = \mathbb{E} \sum_n [\sum_{h \in \mathcal{A}} k_{nh}(N_h^* - 1)/\sqrt{\delta_h}]^2$. For $t > 0$ the displayed formulas give $e^{-2d_{h,t}} = O((\alpha\delta_h t)^{-1})$ and $F_t \leq A_{\text{cert}}/(2\alpha t)$. The draw count gives $A_{\text{cert}} \leq I_* D^2 (N_* - 1)^2/\delta_{\min}$. On a complete b-ary tree, $k_{nh} = |\mathcal{I}(n)|$, so $A_{\text{cert}} = \Theta_2[\sum_h (N_h - 1)/\sqrt{\delta_h}]^2 = O(V^4/(b^2\delta_{\min}))$. The last bound uses $\Theta_2 = \Theta(I^2)$ and $\sum_h (N_h - 1) = b(I-1) + (V-1) = O(V)$. Thus the complete-tree loss rate is $O(V^4/(b^2\alpha\delta_{\min} t))$ whenever its certificate holds. All implied constants are absolute; the positive δ_h remain problem-dependent certificate parameters.

Proof. Each actual-softmax execution weight is a product of factors $1 + x_h$ for correct choices and 1 for wrong choices, divided by products of $N + x_h$. The loss is a finite sum of products of two such execution sums. Clearing denominators gives the stated representation. Coefficient nonnegativity implies nonnegativity at $x \geq 0$. Differentiation therefore gives

$$-\partial_{d_h} F = (1 + x_h) \frac{A\partial_h Q_0 - Q_0 \partial_h A}{Q_0^2} \geq \frac{\delta_h}{(1 + x_h)^2}.$$

The exact update preserves the orthant and increases each relevant margin. The loss decreases along each positive coordinate segment. With $u_t = e^{2d_{h,t}}$, $u_{t+1} \geq u_t \exp(2\alpha\delta_h/u_t) \geq u_t + 2\alpha\delta_h$. Induction gives (61), and substitution into (37) gives (62). Taking the minimum with the already proved monotone loss bound gives its displayed form. All $N_h^* - 1 = 0$ terms vanish for omitted singleton contexts. The finite sum decays at reciprocal-time rate. □

Labelled Boolean height-two illustration. Lemma 12 and Theorem 10 illustrate that the hypotheses of the general convergence results can hold. Use the complete binary height-two shape, and retain the structural pool sizes N_h, head set $\mathcal{H}$, and internal-node count I of that shape. Put $\nu = 2I$ and $Q_0(x) = \prod_{h \in \mathcal{H}} (N_h + x_h)^\nu$. The exact numerator construction and finite verification procedure are in Appendix K.3.

Lemma 12 (Symbolic height-two certificate: Boolean illustration)**.** *Use the zero-vector canonical complete binary height-two model. Each of the following families has explicit constants* $c_h > 0$ *depending only on the chosen gate family:*

1. *Independent fair leaves, independent of an arbitrary exchangeable joint law of the gates on all Boolean truth tables of the given arity.*

2. *Independent uniform AND/OR/XOR gates, independent of an arbitrary exchangeable joint law of the leaves.*

For every loss numerator A in the chosen family, $F = A/Q_0$ and

$$G_h := \nu A - (N_h + x_h)\partial_h A \succeq c_h \frac{Q_0}{(N_h + x_h)^2}. \tag{63}$$

Here $\succeq$ denotes coefficientwise order. Explicitly, let $\mathcal{T}$ be the finite set of gate-multiset component laws in the first family, or the leaf-count component laws in the second, and put $R_h = Q_0/(N_h + x_h)^2$. One may take

$$c_h = \min_{\substack{\tau \in \mathcal{T},\, i \\ [R_h]_i > 0}} \frac{[\nu A_\tau - (N_h + x_h)\partial_h A_\tau]_i}{[R_h]_i} > 0.$$

The positivity and the remaining zero-coefficient comparisons are computer-assisted finite arithmetic statements.

Constants and growth rates. The formula defines each c_h by a finite minimum, with its positivity verified by the finite comparisons below. The coordinate degree is $\nu = \Theta(I)$; tensor interpolation uses $(\nu+1)^r$ evaluations for each component, and coefficient verification uses at most $|\mathcal{T}|r(\nu+1)^r$ comparisons. These counts follow from the coefficient cube in the proof and have absolute prefactors. For this fixed Boolean illustration these structural counts are fixed; the general optimization rates are those of Theorem 9 with $\delta_h = c_h/N_h^3$.

Proof. For each component τ, the numerator construction in (66) and (67) substitutes the actual lower predictions into the root rollout. Summing its squared-residual numerator over the equiprobable component assignments and dividing by twice their count gives A_τ. Each coordinate has degree at most ν, because two copies of the largest true-subtree computation contain at most $2I$ occurrences of any head.

The finite construction evaluates each unnormalized integer numerator on $\{0, \dots, \nu\}^r$. For a polynomial of coordinate degree at most ν, exact interpolation in one variable uses

$$P(x) = \sum_{j=0}^{\nu} \Delta^j P(0) \binom{x}{j}, \qquad \Delta^j P(0) = \sum_{k=0}^{j} (-1)^{j-k} \binom{j}{k} P(k).$$

The forward difference of $\binom{x}{j}$ is $\binom{x}{j-1}$. These polynomials have distinct degrees and nonzero leading coefficients, so they form a basis. Successive differences prove the identity on that basis, hence on every polynomial of the stated degree. Applying it in each coordinate recovers all coefficients by integer and rational operations.

If a_i is an unnormalized coefficient, the corresponding coefficient of the derivative numerator is $g_i = (\nu - i_h)a_i - N_h(i_h + 1)a_{i+e_h}$, with coefficients outside the cube set to zero. Expand R_h by the binomial theorem. If d_τ is the numerator normalization denominator and $[R_h]_i = b_i$, each positive comparison is the exact rational inequality $g_i/d_\tau \geq c_h b_i$. Appendix K.3 specifies the supplied integer verifier and all coefficient comparisons, including strictly positive minimum ratios and nonnegativity when $b_i = 0$. Thus every component satisfies (63). Exchangeability makes the first family a convex mixture of the gate-multiset components and the second a convex mixture of the leaf-count components. Linearity of the population objective preserves every coefficient inequality, proving the claim. □

K.2 Labelled Boolean height-two illustration

The following theorem illustrates that the general drive hypotheses can hold. It concerns only its stated complete binary height-two shape and semantic families. Retain that shape's pool sizes N_h, head set $\mathcal{H}$, and internal-node count I. In the notation of the coefficient certificate, put

$$x_h = e^{d_h} - 1, \qquad \nu = 2I, \qquad A = FQ_0,$$
$$Q_0(x) = \prod_{h \in \mathcal{H}} (N_h + x_h)^\nu, \qquad G_h = \nu A - (N_h + x_h)\partial_h A.$$

For the appropriate finite set $\mathcal{T}$ of gate-multiset or leaf-count component laws and their numerators A_τ, write $R_h = Q_0/(N_h + x_h)^2$ and use the symbolic coefficient minimum

$$c_h = \min_{\substack{\tau\in\mathcal{T},\, i \\ [R_h]_i>0}} \frac{[\nu A_\tau - (N_h + x_h)\partial_h A_\tau]_i}{[R_h]_i}.$$

The certificate and its proof are in Lemma 12 (Appendix K); its symbolic construction is in Appendix K.3.

Theorem 10 (E: certified height-two convergence, Boolean illustration)**.** *Use the zero-vector canonical complete binary height-two model with separate matrices, uniform independent identity injection into $M \geq V$ full-rank tokens satisfying* (41)*, and population GD from zero with any finite $\eta > 0$. Assume one of the following semantic laws:*

1. *Independent fair leaves, independent of an arbitrary exchangeable joint law of the gates on all Boolean truth tables of the given arity.*
2. *Independent uniform AND/OR/XOR gates, independent of any exchangeable joint law of the leaves.*

Let $c_h > 0$ be the constants of Lemma 12 (Appendix K) for the chosen family, and set $\delta_h = c_h/N_h^3$. Then box condition (D) holds with $c = \min_h c_h/N_h^2 > 0$, every margin increases, and

$$F_t \leq \min\left\{F_0, \Theta_2\left(\sum_{h\in\mathcal{H}} \frac{N_h - 1}{\sqrt{1 + 2\alpha\delta_h t}}\right)^2\right\}. \tag{64}$$

Thus the actual full-rollout loss tends to zero. The two semantic families have the separate assumptions stated above.

Constants and growth rates. Here $\delta_h = \Theta(c_h/N_h^3)$ and $c = \Theta(\min_h c_h/N_h^2)$, with absolute implied constants. Writing $\delta_{\min} = \min_h \delta_h$, the rate is $O(\Theta_2[\sum_h (N_h - 1)]^2/(\alpha\delta_{\min} t))$ for $t > 0$, by replacing each denominator in (64) with $\sqrt{2\alpha\delta_{\min} t}$. Its complete-tree structural expression is $O(V^4/(b^2\alpha\delta_{\min} t))$ by Theorem 9 (Appendix K); this illustration uses its fixed binary height-two shape and certified component laws.

Proof. Exchangeability makes distinct placements of a given gate multiset uniform. Thus the first law is a convex mixture of the certified gate-multiset component laws. The second is a convex mixture of the certified leaf-count laws, since leaf exchangeability makes arrangements of a fixed count uniform. Linearity of the objective preserves all coefficient inequalities. For the resulting numerator A and G_h, differentiation gives

$$-\partial_{d_h} F = \frac{(1 + x_h)G_h}{(N_h + x_h)Q_0} \geq c_h \frac{1 + x_h}{(N_h + x_h)^3} \geq \frac{c_h}{N_h^3(1 + x_h)^2}.$$

The last inequality follows from $N_h + x_h \leq N_h(1 + x_h)$. Also $\gamma_h = x_h/(N_h + x_h)$, whence $-\partial_{\gamma_h} F = (N_h + x_h)G_h/(N_h Q_0) \geq c_h/[N_h(N_h + x_h)] = (c_h/N_h^2)(1 - \gamma_h)$. This proves (D) on the whole box by continuity at its boundary. The margin induction in Theorem 9 (Appendix K) therefore applies with the displayed δ_h. Each head occurs $|\mathcal{I}(n)|$ times in the true computation of output n. Hence $k_{nh} = |\mathcal{I}(n)|$ and $\sum_n k_{nh}^2 = \Theta_2$. Factoring the common subtree count out of the inner sum in (62) (Appendix K) gives (64). □

K.3 Exact finite certificate construction

This subsection supplies the finite arithmetic component of Lemma 12. All theoretical constants remain symbolic. For the complete binary height-two tree, label leaves a, b, c, d, lower internal nodes u, v, and root o. Let T_u, T_v, T_o be their truth tensors and let N_1, N_2, N_f be the structural pool sizes. Put

$$D_0 = \prod_{h\in\mathcal{H}} (N_h + x_h), \qquad Q_0 = D_0^\nu, \qquad \nu = 2I. \tag{65}$$

A bracketed sum of truth tensors is evaluated on the two vector arguments following it. The exact rollout numerators are

$$\begin{aligned}
P_u &= [(1+x_f)T_u + T_v + T_o]((1+x_1)e_a + e_c, (1+x_2)e_b + e_d),\\
P_v &= [T_u + (1+x_f)T_v + T_o](e_a + (1+x_1)e_c, e_b + (1+x_2)e_d),\\
R_1 &= D_0(e_a + e_c) + (1+x_1)P_u,\\
R_2 &= D_0(e_b + e_d) + (1+x_2)P_v,\\
P_o &= [T_u + T_v + (1+x_f)T_o](R_1, R_2), \qquad (66)\\
W_0 &= D_0^{\nu-2}\big(\|D_0 e_{y_u} - P_u\|^2 + \|D_0 e_{y_v} - P_v\|^2\big)\\
&\quad + \|D_0^I e_{y_o} - P_o\|^2. \qquad (67)
\end{aligned}$$

Here e_a, e_b, e_c, e_d denote leaf value one-hots. The prediction denominators are D_0, D_0, D_0^I, and the literal per-instance loss is $W_0/(2Q_0)$. For each component law τ, let Ω_τ be its equally weighted semantic assignments. Then

$$A_\tau = \frac{1}{2|\Omega_\tau|}\sum_{\omega\in\Omega_\tau} W_0(\omega).$$

The first family uses uniform distinct placements of each Boolean-gate multiset and independent fair leaves. The second uses independent uniform AND/OR/XOR gates and uniform placements of each fixed count of one-valued leaves. These define the finite sets $\mathcal{T}$ and Ω_τ completely.

For coordinate degree at most ν, values on $\{0,\dots,\nu\}^{|\mathcal{H}|}$ determine the coefficients exactly by the Newton interpolation identity proved in Lemma 12. If $a_{\tau,i}$ is the integer coefficient before division by $d_\tau = 2|\Omega_\tau|$, define

$$g_{\tau,h,i} = (\nu - i_h)a_{\tau,i} - N_h(i_h+1)a_{\tau,i+e_h}, \qquad b_{h,i} = [Q_0/(N_h + x_h)^2]_i.$$

Here $[\cdot]_i$ is the coefficient of $x^i = \prod_h x_h^{i_h}$. Writing $A_\tau = d_\tau^{-1}\sum_i a_{\tau,i}x^i$, one has $x_h\partial_h A_\tau = d_\tau^{-1}\sum_i i_h a_{\tau,i}x^i$ and, after the index shift $j = i - e_h$, $\partial_h A_\tau = d_\tau^{-1}\sum_j (j_h+1)a_{\tau,j+e_h}x^j$. Hence $g_{\tau,h,i}/d_\tau = [\nu A_\tau - (N_h + x_h)\partial_h A_\tau]_i$, the coefficient compared in Lemma 12. Coefficients outside the cube are zero. The supplied exact verifier checks every $g_{\tau,h,i} \ge 0$ and checks strict positivity whenever $b_{h,i} > 0$. Consequently the finite minimum

$$c_h = \min_{\tau,i:\, b_{h,i}>0} \frac{g_{\tau,h,i}}{d_\tau b_{h,i}}$$

is strictly positive and proves every coefficient inequality in (63). This is the computer-assisted step. It uses exact integers and rational comparisons, with no floating-point positivity tolerance.

The two coefficient arrays, their reconstruction code, the positivity verifier, and its output are supplied in `theory_checks/certificates/`. The verifier output records the component counts, comparison counts, symbolic-minimum evaluations and file hashes. The manuscript uses only positivity and the displayed symbolic definitions. This certificate is distinct from the sampled generalization sanity check.

L Proof of the Generalization Bound

The argument uses the final-embedding loss of Eq. (9), with its factor $1/2$. All immutable blocks and all leaf values are copied exactly, so the test loss is the sum of the half-squared errors of the internal value coefficients. We give the argument for an individual test instance and identity injection; averaging therefore gives the same bound for any test distribution satisfying the stated bounds.

Score gaps and witness coverage. Let $Z_h = B^\top B^h B$ be the score table of head h, with key identity as its row index and target identity as its column index. Put

$$g_{jk}^h = Z_h[k,k] - Z_h[j,k], \qquad m(Z) = \min_{h\in\mathcal{A}}\min_{j\ne k} g_{jk}^h.$$

The minimum is finite and attained. In a child pool, distinct slot-h candidates have distinct parents, and the correct candidate has the target as its parent. In a function pool, the correct key is the

target itself. Thus every correct-versus-wrong test logit gap of a head in $\mathcal{A}$ is one of these g^h_{jk}. The assumptions that identities come from the same public set and that the test nontrivial heads belong to $\mathcal{A}$ ensure that all nontrivial test contexts are covered. A singleton context is exact without any gap assumption.

For completeness, a head-h null witness is a pair (n, v) where n is a queried internal target and v is a wrong candidate whose read is identically zero. For a child head this means v occupies slot h and $\mathrm{ht}(v) \geq \mathrm{ht}(n)$; for the function head it means v is a leaf. Its wrong key is the identity of $p(v)$ for a child head and of v for the function head. Write W_h^{wit} for the expected number of these pairs under the training shape law and $W_{\min} = \min_{h\in\mathcal{A}} W_h^{\mathrm{wit}} > 0$.

From training loss to every score gap. The score-table barrier in Corollary 2 uses the canonical model, uniform independent injection, bounded candidate pools, positive witness counts, and supervision of every internal output; its proof requires no rank condition on B and gives

$$\mathcal{L}_{\mathrm{train}}(\theta) \geq \frac{W_{\min}}{2M(M-1)(e^{m(Z)} + N_* - 1)^2}. \tag{68}$$

Define

$$u = \sqrt{\frac{2M(M-1)\varepsilon}{W_{\min}}}, \qquad a = N_* - 1.$$

Combining (68) with $\mathcal{L}_{\mathrm{train}} \leq \varepsilon$ gives $e^{m(Z)} + a \geq u^{-1}$. By (17), $au \leq 1/2$, so $u^{-1} - a$ is positive and

$$e^{-m(Z)} \leq \frac{u}{1-au} \leq 2u, \qquad e^{-2m(Z)} \leq \frac{8M(M-1)\varepsilon}{W_{\min}}. \tag{69}$$

The threshold and inversion use population training loss. Replacing it by an empirical estimate would require a separate sampling argument.

Wrong mass of one draw. Put $N' = \max\{2, V'\}$, which bounds every test pool. In a nontrivial pool of size $N \leq N'$, let k be the target identity and let J be its wrong key identities. Every gap is at least $m(Z)$, and softmax column-shift invariance gives

$$\Pr(\text{wrong draw}) = \frac{\sum_{j\in J} e^{-g^h_{jk}}}{1 + \sum_{j\in J} e^{-g^h_{jk}}} \leq \sum_{j\in J} e^{-g^h_{jk}} \leq (N-1)e^{-m(Z)} \leq (N'-1)e^{-m(Z)}. \tag{70}$$

The same upper bound holds for singleton pools because their wrong mass is zero.

Execution error and the test loss. Fix an internal test node n. By Lemma 5, its soft value coefficients satisfy $z_n = \mathbb{E}[e_{Y_n} \mid \mathcal{G}', \iota]$, where the hard execution Y_n is constructed by independently drawing sources at each semantic occurrence and recursively unfolding earlier written states. Null results have vector representation zero. Consequently

$$\epsilon_n = 1 - z_{n,y_n} = \Pr(Y_n \neq y_n), \qquad \tfrac{1}{2}\|z_n - e_{y_n}\|_2^2 \leq \epsilon_n^2.$$

If all draws in the true subtree execution choose their correct sources, the structural height schedule makes every child available, the function code is correct, and calibration returns the true value. To bound the complementary event, expose draws until the first incorrect choice. Conditional on all preceding draws being correct, each queried internal node v lies in the true subtree of n, and its queries consist of one function read and one read for each existing child slot. Their pool sizes are V' and N'_h, respectively. Along this all-correct execution each such query occurs exactly once. Applying (70) to the mutually exclusive possible locations of the first error gives

$$\epsilon_n \leq \Pr(\text{some draw is wrong}) \leq \min\{1, A_n e^{-m(Z)}\},$$

where

$$A_n = \sum_{v\in\mathcal{I}'(n)} \left[V' - 1 + \sum_{h=1}^{a_v} (N'_h - 1) \right].$$

Singleton reads contribute zero. The number of draws after a wrong choice is immaterial, because the event has already been counted. This argument does not assume independence of stored soft

states or of semantic variables. Independence is only the conditional independence in the unfolded execution of Lemma 5; the first-error union bound also holds directly for adaptive exposure of the draws. Summing over the I' internal nodes and applying (69) proves

$$\mathcal{L}_{\text{test}}(\mathcal{G}';\theta) = \frac{1}{2}\sum_{n\in\mathcal{I}'} \|z_n - e_{y_n}\|_2^2 \le \sum_{n\in\mathcal{I}'} \min\{1, A_n^2 e^{-2m(Z)}\}$$
$$\le \min\left\{I', \frac{8\Gamma(\mathcal{G}')M(M-1)}{W_{\min}}\,\varepsilon\right\}.$$

All quantities here refer to the same finite attention matrices used in training and testing. The proof requires neither an equicorrelated Gram matrix nor a particular training trajectory.

Subtree counts and growth rates. Write $J_n = |\mathcal{I}'(n)|$ and $\Theta_2(\mathcal{G}') = \sum_n J_n^2$. Since child slots partition all nonroot nodes, $\sum_h N'_h = V' - 1$. Consequently $A_n \le 2(V'-1)J_n$ and

$$\Gamma(\mathcal{G}') \le 4(V'-1)^2\Theta_2(\mathcal{G}') \le 4(V'-1)^2 I'^3.$$

Thus the bound is polynomial in the actual test size even for irregular trees. It also recovers the coarser execution count: if each all-correct execution has at most D' draws and all pools have size at most N', then $A_n \le D'(N'-1)$ and $\Gamma \le I'D'^2(N'-1)^2$.

For a complete b-ary tree of positive height K', $I' = (V'-1)/b$ and every child pool has I' candidates. Each internal node therefore contributes $b(I'-1) + (V'-1) = 2(V'-1) - b$ to A_n. Hence

$$\Gamma(\mathcal{G}') = [2(V'-1) - b]^2 \sum_{\ell=1}^{K'} b^{K'-\ell}\left(\frac{b^\ell - 1}{b-1}\right)^2.$$

By (30), $I'^2 \le \Theta_2 \le bI'^2/(b-1) \le 2I'^2$. Also $V' - 1 \le 2(V'-1) - b \le 2(V'-1)$. These inequalities give, uniformly for $b \ge 2$,

$$C_{\text{gen}} := \frac{8\Gamma(\mathcal{G}')M(M-1)}{W_{\min}} = \Theta\left(\frac{V'^4 M^2}{b^2 W_{\min}}\right).$$

For fixed b, this is fourth-power rather than fifth-power growth in test size. Since $V' = \Theta(b^{K'})$, it is still exponential in test depth. This is a sufficient loss-transfer bound, not a lower bound on the model's error. For general trees retain the shape-dependent $\Gamma(\mathcal{G}')$. The training threshold remains $\varepsilon_{\text{gen}} = \Theta(W_{\min}/(M^2 N_*^2))$.

Training budget. If the hypotheses of Theorem 3 also hold and $I' > 0$, choose $\varepsilon = \min\{\varepsilon_{\text{gen}}, \delta/C_{\text{gen}}\}$ for any target $\delta > 0$. Then

$$T_1 + \left\lceil \frac{\rho}{\alpha(1 - \alpha L_d/2)\varepsilon} \right\rceil$$

iterations suffice for test loss at most δ. This substitution preserves the theorem's explicit dependence on the drive floor and on the dictionary conditioning through α. A test tree with no internal node has loss zero directly.

Root-answer evaluation. The same argument applies to the root o alone, without summing over outputs:

$$\mathcal{L}_{\text{root}} := \tfrac{1}{2}\|z_o - e_{y_o}\|_2^2 \le \min\left\{1, \frac{8A_o^2 M(M-1)}{W_{\min}}\,\varepsilon\right\}.$$

For nearest-symbol decoding $\widehat{y}_o \in \arg\max_{a\in[q]} z_{o,a}$, an incorrect answer implies $z_{o,a} \ge z_{o,y_o}$ for some $a \ne y_o$. Minimizing $(1 - z_{o,y_o})^2 + z_{o,a}^2$ under this constraint gives $1/2$, so

$$\mathbf{1}\{\widehat{y}_o \ne y_o\} \le 4\mathcal{L}_{\text{root}}.$$

In particular, $\mathcal{L}_{\text{root}} < 1/4$ ensures the correct root answer, and averaging gives an error-probability bound. This is an evaluation consequence of training with all internal outputs supervised. It does not assert convergence from root-only supervision, or convert language-model teacher-forced cross-entropy into free-running answer accuracy.

Capacity of the identity dictionary. The argument above uses the score tables $Z_h = B^\top B^h B$ and the barrier (68), which holds for arbitrary finite score tables; it uses no rank condition on B. The hypothesis $M \geq V'$ is needed because test identities are distinct public tokens. Full column rank of B, and hence $d \geq M$, is required by the exact interpolation of Theorem 1 and by the optimization results, including Theorems 2 and 3. Small training loss itself is available with more identities than dimensions. If the public tokens have unit norm and pairwise inner products at most $\mu < 1$, the choice $B^h = cI_d$ with $c > 0$ gives $Z_h = cB^\top B$, so every gap satisfies $g^h_{jk} = c(1-b_j^\top b_k) \geq c(1-\mu)$. The execution bound of Lemma 5 then gives $\mathcal{L}_{\mathrm{train}} \leq I_* D^2 (N_* - 1)^2 e^{-2c(1-\mu)}$, which tends to zero as $c \to \infty$, so the bound of this appendix applies. Whether gradient descent from zero reaches such parameters when $M > d$ is not addressed by our results.