Understanding the Out-of-Distribution Generalization of Chain-of-Thought Reasoning in LLMs

Nuo-Jing Liang[1] **Xiao-Tong Yuan**[1]* **Pan Zhou**[2]
[1]School of Intelligence Science and Technology, Nanjing University [2]Singapore Management University
nuojingliang@smail.nju.edu.cn, xtyuan@nju.edu.cn, panzhou@smu.edu.sg

Abstract

Chain-of-Thought (CoT) reasoning substantially improves Large Language Models (LLMs), yet its out-of-distribution (OOD) generalization mechanism remains theoretically underexplored. We study CoT as compositional OOD generalization: the target task lacks complete reasoning trajectories or direct input-output pairs, while training provides only in-distribution (ID) atomic step data. We show that CoT generalizes by repeatedly applying a shared single-step predictor to compose reusable ID transitions. Theoretically, let T denote the number of reasoning steps in the target task, n the number of training samples per sub-task, and k the number of output tokens per step; then the OOD task risk is controlled by the sum of the ID sub-task risks, scaled by a coefficient of order $\mathcal{O}(k)$ that is independent of T, although the summed risk itself may still grow with T. A Rademacher-complexity analysis further shows that the estimation term is of order $\mathcal{O}(\sqrt{T/n})$, rather than linear in T, due to predictor sharing under a uniform capacity control on the shared class. For autoregressive Next-Token Prediction, the final-task risk is bounded by the best-in-class step-level risk scaled by the same coefficient of order $\mathcal{O}(k)$, plus an estimation term of order $\widetilde{\mathcal{O}}\big(k\sqrt{T/n}\big)$. A parity-coupled construction further shows that, for the sub-task empirical risk minimizer, the expected final-task risk is at least the best in-class risk plus $\Omega(\sqrt{T/n})$ when $T \leq n$. Experiments on Atom-Task and GSM8K support the theory: CoT enables short-to-long and compositional generalization, while explicit state tracking is crucial.

Our code is available at https://github.com/maxliang114514/CoT-OOD.

1 Introduction

Chain-of-Thought (CoT) reasoning has become a central mechanism for improving the reasoning ability of Large Language Models (LLMs). By prompting or training models to generate intermediate steps, CoT substantially improves mathematical and logical reasoning [1–8]. Recent reasoning-oriented systems further treat step generation as an explicit training signal [9–11], shifting LLM reasoning from direct pattern matching toward structured stepwise computation [12]. Despite this empirical success, a fundamental theoretical question remains open: *why can CoT-trained models generalize to complex reasoning tasks that are not directly observed during training?*

Existing theoretical studies provide important insights into the expressivity and computational role of CoT. Some works show that CoT can increase the effective depth of Transformers or enable serial computation [13–15], while others analyze task decomposition, in-context learning, or localized reasoning settings [16–18]. However, most existing analyses are either qualitative, restricted to narrow or two-step settings, or do not provide explicit generalization bounds for the autoregressive Next-Token Prediction (NTP) regime commonly used by modern large-scale LLMs. As a result,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the out-of-distribution (OOD) generalization mechanism of multi-step CoT remains theoretically underexplored.

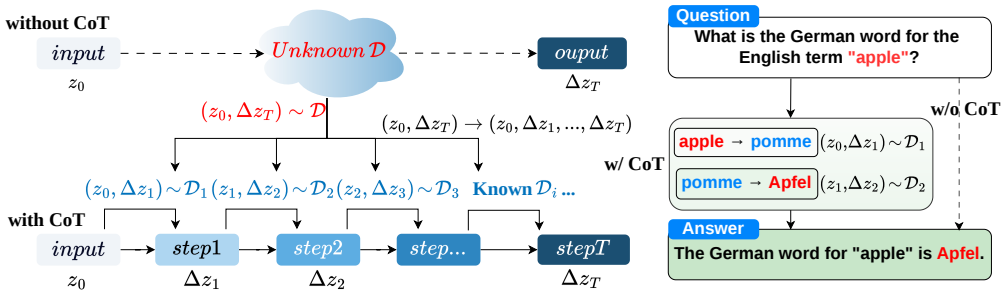


Figure 1: CoT as compositional OOD generalization. Left: without CoT, the model must learn a direct mapping from input z_0 to final answer Δz_T, whose target distribution may be unavailable during training. With CoT, the same predictor is repeatedly applied to in-distribution step distributions, mapping each state z_{i-1} to update Δz_i. Right: a two-hop translation example solves an unseen English-to-German query by composing learned English-to-French and French-to-German steps.

This paper studies the mechanism of CoT through compositional OOD generalization. We consider a target task requiring a complex multi-step mapping from an input state z_0 to a final answer Δz_T, while neither complete reasoning trajectories $(z_0, \Delta z_1, \ldots, \Delta z_{T-1}, \Delta z_T)$ nor direct input-output pairs $(z_0, \Delta z_T)$ are available during training. Thus, the target task's composition is OOD relative to the observed atomic transitions. We refer to this setting as compositional OOD; Section 3.2 separately treats mismatched training and target step distributions (Corollary 4). Instead, the model has access to in-distribution (ID) atomic step data: samples $(z_{i-1}, \Delta z_i)$, where z_{i-1} is the reasoning state before step i and Δz_i is the next incremental output. Here, T denotes the total number of generation steps: the first $T-1$ are intermediate reasoning updates, and Δz_T is the final answer.

Under this view, CoT does not require the model to learn the entire OOD mapping at once from scratch. Instead, it repeatedly applies a shared single-step predictor. At each step, the model reads the current state z_{i-1}, generates Δz_i containing k new tokens, and updates the state by appending these tokens while keeping a fixed context length m. CoT therefore converts a complex target mapping into a composition of reusable ID atomic transitions (Fig. 1). This motivates our central question:

When a model learns only atomic step transitions, can repeated CoT composition generalize to unseen multi-step OOD tasks, and how does the resulting error scale with the reasoning depth T?

Contributions We answer this question by developing a generalization framework for CoT-based task decomposition. First, we prove that the risk of the complex OOD task can be upper-bounded by the sum of risks over ID sub-tasks, scaled by a coefficient of order $\mathcal{O}(k)$ that is independent of T. The coefficient results from deterministic argmax decoding together with the bounded sequence loss and the local error floor $(\log 2)/k$ that a hard-decoding (argmax) error induces. CoT thus avoids fitting the unseen composite mapping directly and composes local transitions at test time.

Second, we derive a Rademacher-complexity bound for the multi-step setting. A naive analysis that treats each reasoning step independently would yield a linear dependence on T. In contrast, because CoT applies the same predictor across all steps, we can analyze the combined step-level sample set and obtain a sublinear generalization-gap term scaling as $\mathcal{O}(\sqrt{T/n})$, where n is the number of training samples per atomic step distribution.

Third, we specialize the analysis to the NTP setting of autoregressive LLMs. We directly analyze the final k target tokens of each atomic sequence and take the $N = nT$ complete sequences as the sampling units. The resulting final-task risk bound keeps the best in-class step-level risk explicit and adds an estimation term of order $\widetilde{\mathcal{O}}\big(k\sqrt{T/n}\big)$. Appendix B.10 shows that the $\sqrt{T/n}$ dependence cannot be improved for the sub-task empirical risk minimizer: a construction with $k = 1$, in which the per-step errors are coupled through the XOR accumulated over the chain, makes this minimizer incur an excess of $\Omega(\sqrt{T/n})$ when $T \le n$, and carries that excess through to the final task.

We further validate our theory through controlled experiments. We introduce Atom-Task, a symbolic compositional benchmark involving multi-step sequence operations, to isolate the effect of CoT on OOD compositional generalization. Experiments show that CoT-trained models generalize from shorter or constrained training chains to longer unseen compositions, whereas models trained without intermediate reasoning fail to capture the underlying compositional structure. Ablation results further show that explicitly tracking the current reasoning state is crucial, matching our theoretical requirement that each state z_{i-1} contains sufficient information for the next atomic transition. Additional experiments on GSM8K, a benchmark of grade-school math word problems, support the same conclusion in real-world mathematical reasoning.

2 Preliminaries

Problem Setting We consider an autoregressive predictor with a context window size of m. Let $\mathcal{V}$ be a vocabulary (a set of tokens), with $n_v = |\mathcal{V}|$ denoting its size, and let $\mathcal{V}^m$ denote the set of all token sequences of length m formed by tokens from $\mathcal{V}$. The predictor $\psi : \mathcal{V}^m \to \mathbb{R}^{n_v \times k}$ ($m \gg k$) maps an input sequence to probability distributions over k output positions, and belongs to a hypothesis class Ψ. Losses are computed on these soft predictions; for autoregressive generation, discrete tokens are obtained via argmax, denoted $\mathrm{hard}(\psi(\cdot)) \in \mathcal{V}^k$.

Under the CoT paradigm, the model does not predict the final answer directly via $\psi(z_0)$, but generates a chain of T reasoning steps. We write $\widehat{\cdot}$ for model-generated quantities and reserve the unhatted symbol for the ground truth. At step i, the incremental output is $\widehat{\Delta z}_i = \mathrm{hard}(\psi(\widehat{z}_{i-1}))$, and the state updates by appending these k tokens while discarding the first k (sliding window):

$$\widehat{z}_i = \widehat{z}_{i-1}[k{+}1 : m] \oplus \widehat{\Delta z}_i \quad (\oplus \text{ denotes sequence concatenation}), \quad \widehat{\Delta z}_T = \mathrm{hard}(\psi(\widehat{z}_{T-1})),$$

thus maintaining the sequence length m for all intermediate states, with $\widehat{z}_0 = z_0$; the true state follows the same recursion with Δz_i in place of $\widehat{\Delta z_i}$, i.e., $z_i = z_{i-1}[k+1 : m] \oplus \Delta z_i$.

For a given complex task, any instance $(z_0, \Delta z_T)$ with $z_0 \in \mathcal{V}^m$ and $\Delta z_T \in \mathcal{V}^k$ satisfies $(z_0, \Delta z_T) \sim \mathcal{D}$. There exists a ground-truth mapping ψ_* (not necessarily within Ψ) such that $\mathrm{hard}(\psi_*(z_{i-1})) = \Delta z_i$ for all i, implying ψ_* generates the CoT sequence $(\Delta z_1, \ldots, \Delta z_T)$ from z_0 to yield the target output Δz_T.

Define the state transition $\bar{\psi} : \mathcal{V}^m \to \mathcal{V}^m$ as $\bar{\psi}(z) := z[k{+}1 : m] \oplus \mathrm{hard}(\psi(z))$, representing the input state for the next step (note $\bar{\psi}$ uses hard predictions for transitions). To distinguish from ψ, let $F_\psi : \mathcal{V}^m \to \mathbb{R}^{n_v \times k}$ denote the composite CoT mapping $F_\psi(z_0) := \psi(\bar{\psi}^{(T-1)}(z_0))$, which encapsulates the T-step reasoning process (note F_ψ applies ψ to the final state, outputting soft predictions for loss computation). Our learning goal is to minimize the population risk:

$$R_{\mathcal{D}}(F_\psi) = \int \mathfrak{L}(F_\psi(z_0), \Delta z_T)\, dP(z_0, \Delta z_T) = \int \mathfrak{L}(\psi(\bar{\psi}^{(T-1)}(z_0)), \Delta z_T)\, dP(z_0, \Delta z_T),$$

where P is the probability measure associated with $\mathcal{D}$ and $\mathfrak{L} : \mathbb{R}^{n_v \times k} \times \mathbb{R}^{n_v \times k} \to \mathbb{R}$ is the sequence-level loss function that compares the model's soft predictions (in $\mathbb{R}^{n_v \times k}$) with the ground-truth output Δz_T (converted from discrete tokens to one-hot vectors in $\mathbb{R}^{n_v \times k}$); the formal definition is provided in Section 3. This aims to identify a $\psi \in \Psi$ approximating ψ_*, thereby generating the CoT sequence from z_0 to yield Δz_T.

Notations $\bar{\psi}^{(T)}$ denotes the T-fold composition $\bar{\psi} \circ \cdots \circ \bar{\psi}$ (T times in total). The final prediction is extracted via $\psi(\bar{\psi}^{(T-1)}(z_0))$. For a sequence z, $z[a : b]$ denotes the subsequence from index a to b (inclusive), and $(z)_j$ denotes its j-th component (or token). The operator $\oplus$ denotes the concatenation of two sequences. Other key notations are summarized in Table 4 in Appendix B.

3 Theoretical Analysis: Task Decomposition for OOD Tasks

We analyze the generalization error of CoT reasoning by decomposing the risk of the complex OOD task into per-step risks (Section 3.1) and then bounding the generalization gap (Section 3.2).

Assumption 1 (Bounded and Lipschitz Loss Function)**.** For a single token represented as a probability vector in $\mathbb{R}^{n_v}$, the token-level loss function $\ell(\cdot, \cdot)$ is defined as the ϵ_ℓ-clipped cross-entropy loss

with a fixed clipping threshold $\epsilon_\ell \in (0, 1/4)$, i.e., $\ell(p, y) := -\log \max\{p_y,\ \epsilon_\ell\}$, where p denotes the predicted probability vector and y the ground-truth token. It is upper-bounded by B_ℓ, i.e., $\ell(p, y) \le B_\ell$, and is G_ℓ-Lipschitz continuous with respect to the ℓ_2-norm of the predicted probability vector, i.e., $|\ell(\hat{y}_1, y) - \ell(\hat{y}_2, y)| \le G_\ell \|\hat{y}_1 - \hat{y}_2\|_2$ for all probability vectors $\hat{y}_1, \hat{y}_2 \in \Delta_{n_v - 1} := \{p \in [0, 1]^{n_v} : \sum_{a=1}^{n_v} p_a = 1\}$ and ground-truth token y. Without loss of generality, assume $G_\ell \ge 1$.

Remark 1. Assumption 1 is standard in learning theory [19–22]. Both requirements follow directly from clipping, which lower-bounds the probability assigned to the true token by ϵ_ℓ; both conditions are therefore satisfied by construction.

Definition 1 (Sequence Loss)**.** For a predicted probability distribution $\hat{y} \in \mathbb{R}^{n_v \times k}$ and a ground-truth token sequence Δz (represented as one-hot vectors in $\mathbb{R}^{n_v \times k}$), the sequence loss is defined as the average of the token-wise loss ℓ over these k generated tokens, where $\hat{y}_j$ denotes the probability vector at output position j (the j-th column of $\hat{y}$) [23]:

$$\mathfrak{L}(\hat{y}, \Delta z) := \frac{1}{k} \sum\nolimits_{j=1}^{k} \ell(\hat{y}_j, (\Delta z)_j)$$

3.1 Generalization Error Decomposition via CoT

For $(z_0, \Delta z_T) \sim \mathcal{D}$, let $\mathcal{D}'$ denote the joint distribution over the reasoning sequence $(z_0, \Delta z_1, \ldots, \Delta z_T)$, and let $\mathcal{D}_i$ denote the marginal law of $(z_{i-1}, \Delta z_i)$ under $\mathcal{D}'$ (a state on the ground-truth trajectory is called an oracle state). Lemma 7 in Appendix B.2 shows that the step-level loss of an erroneous hard-decoded block is at least $(\log 2)/k$, a floor that is independent of the reasoning depth T.

Theorem 1 (Generalization Error Decomposition)**.** *Under Assumption 1, we have:*

$$R_{\mathcal{D}}(F_\psi) \le \frac{B_\ell k}{\log 2} \sum\nolimits_{i=1}^{T} \int \mathfrak{L}(\psi(z_{i-1}), \Delta z_i) d\mathcal{D}_i(z_{i-1}, \Delta z_i),$$

where $(\log 2)/k$ is the local error floor of Lemma 7.

Remark 2 (Implications of Theorem 1). Theorem 1 reveals the fundamental principle underlying CoT's OOD generalization: *the generalization risk of a complex OOD task is dominated by the sum of per-step in-distribution risks*, scaled by an $\mathcal{O}(k)$ coefficient that is independent of T. This is qualitatively different from direct OOD learning, where one would attempt to bound the risk of the composite mapping F_ψ directly—without decomposition, not only is the mapping F_ψ inherently more complex, but more critically, we lack the direct samples $(z_0, \Delta z_T) \sim \mathcal{D}$, so the composite risk cannot be estimated from target data alone.

Theorem 1 bounds the target risk by the summed step-level risks, and the additive dependence on those risks is unavoidable in general. A step-level error can propagate along the chain through the state update rather than being averaged out, so the per-step risks accumulate in the final answer.

Remark 3 (Tightness of Theorem 1). Appendix B.9 constructs, for $k = 1$, a binary sliding-window task with a shared single-step predictor ψ and a bounded sequence loss, for which

$$R_{\mathcal{D}}(F_\psi) = \Omega\left(\sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi) \right).$$

Consequently, the additive form of Theorem 1 is tight up to its depth-independent coefficient: when every step-level error reaches the final answer, the linear growth in T of the summed risk is attained, and no argument that charges only per-step risks can remove it.

3.2 Bounds via Rademacher Complexity

Let $S_i = \{(z_{i-1}^{(j)}, \Delta z_i^{(j)})\}_{j=1}^{n}$ be n independent atomic pairs drawn from $\mathcal{D}_i$, with $S_0 = \bigcup_{i=1}^{T} S_i$. All nT atomic examples are mutually independent, with the n examples in S_i drawn from $\mathcal{D}_i$; each comes from some source realizing the same step-level skill distribution and need not lie on a common trajectory. Define $f_\psi(z, \Delta z) := \mathfrak{L}(\psi(z), \Delta z)$ and $\mathcal{F} := \{f_\psi : \psi \in \Psi\}$. To bound the generalization gap we use Rademacher complexity, writing $R_{\mathcal{D}_i}(\psi)$ and $\widehat{R}_{S_i}(\psi)$ for the associated population and empirical risks (Appendix B.4). Applying Rademacher bounds to the T steps separately makes the resulting gap grow linearly in T. Since the same learner ψ processes samples from all T

distributions, analyzing the combined set S_0 under a uniform capacity control makes both the deviation term and the complexity term (from $\sum_{i=1}^{T} \widehat{\mathfrak{R}}_{S_i}(\mathcal{F})$ to $T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})$) improve from $\mathcal{O}(T)$ to $\mathcal{O}(\sqrt{T})$—analogous to multi-task learning, where sharing a hypothesis across related tasks improves generalization [24]. Extending the classical symmetrization lemma [19] to multiple distributions (Lemma 8 in Appendix B.4) gives:

Theorem 2 (Rademacher Complexity Bound)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over $\psi \in \Psi$, we have:*

$$\left|\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi) - \widehat{R}_{S_i}(\psi))\right| \le 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) + 3B_\ell\sqrt{\frac{2T\log(4/\delta)}{n}},$$

where $\widehat{\mathfrak{R}}_{S_0}(\cdot)$ denotes the empirical Rademacher complexity on S_0, $\mathcal{F} = \mathfrak{L} \circ \Psi$ is the induced bounded loss class, and B_ℓ is the bound of the sequence loss $\mathfrak{L}$ according to Assumption 1.

Remark 4. Sharing a single learner ψ across all T reasoning steps, under a uniform capacity control on the shared class, is what permits treating the T step-level datasets as one combined sample S_0 of size nT: the same predictor is fitted to all steps, so the estimation dependence improves from T to $\sqrt{T}$.

The following corollary records the resulting guarantee, first for a fixed predictor and then, after the ERM comparison, for the empirical risk minimizer; both proofs are in Appendix B.5.

Corollary 3 (Generalization Bound for Complex OOD Tasks)**.** *Under Assumption 1, with probability at least $1-\delta$, for every $\psi \in \Psi$,*

$$R_{\mathcal{D}}(F_\psi) \le \frac{B_\ell k}{\log 2}\Big[\sum_{i=1}^{T}\widehat{R}_{S_i}(\psi) + 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) + 3B_\ell\sqrt{\frac{2T\log(4/\delta)}{n}}\Big],$$

and for the sub-task ERM $\widehat{\psi} \in \arg\min_{\psi\in\Psi}\sum_{i=1}^{T}\widehat{R}_{S_i}(\psi)$, with $\psi^\star \in \arg\min_{\psi\in\Psi}\sum_{i=1}^{T}R_{\mathcal{D}_i}(\psi)$,

$$R_{\mathcal{D}}(F_{\widehat{\psi}}) \le \frac{B_\ell k}{\log 2}\Big[\sum_{i=1}^{T}R_{\mathcal{D}_i}(\psi^\star) + 4T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) + 6B_\ell\sqrt{\frac{2T\log(4/\delta)}{n}}\Big].$$

Remark 5 (Tightness of Corollary 3). The estimation term of Theorem 2 is order-tight, and its rate is inherited by Corollary 3: Appendix B.10 constructs a binary compositional task and a two-element class for which the sub-task ERM $\widehat{\psi}$ of Corollary 3 satisfies

$$\mathbb{E}_S\left[\sum_{i=1}^{T}R_{\mathcal{D}_i}(\widehat{\psi}) - \min_{\psi\in\Psi}\Big\{\sum_{i=1}^{T}R_{\mathcal{D}_i}(\psi)\Big\}\right] = \Omega\left(\min\left\{1, \sqrt{\frac{T}{n}}\right\}\right),$$

i.e., its *atomic excess*, the gap between its summed step-level risk and the best in-class value of that sum, is of that order; and in that instance the atomic excess coincides with the excess on the final task, so the same bound holds for $\mathbb{E}_S[R_{\mathcal{D}}(F_{\widehat{\psi}}) - \min_{\psi\in\Psi}R_{\mathcal{D}}(F_\psi)]$. In particular it is $\Omega(\sqrt{T/n})$ when $T \le n$. Two features of the construction are responsible: the per-step separation is amplified by a factor of T at the final token, and the best-in-class per-step risk is $\Theta(1)$, which makes the excess the identifiable quantity.

The preceding analysis addresses compositional OOD generalization: the target task is out of distribution, whereas the step-level distributions $\mathcal{D}_i$ are not. When the step-level data are themselves drawn from shifted distributions Q_i, the decomposition carries over with one additional per-step term.

Corollary 4 (Atomic Distribution Shift)**.** *Let Q_i be the training distribution at step i, let $\mathcal{D}_i$ be the corresponding target oracle-state marginal, and let $S_i \sim Q_i^n$. Define*

$$\mathrm{disc}_{\mathcal{F}}(\mathcal{D}_i, Q_i) := \sup_{\psi\in\Psi}\left|R_{\mathcal{D}_i}(\psi) - R_{Q_i}(\psi)\right|.$$

Then, with probability at least $1-\delta$, for every $\psi \in \Psi$,

$$R_{\mathcal{D}}(F_\psi) \le \frac{B_\ell k}{\log 2}\left[\sum_{i=1}^{T}\widehat{R}_{S_i}(\psi) + 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) + 3B_\ell\sqrt{\frac{2T\log(4/\delta)}{n}} + \sum_{i=1}^{T}\mathrm{disc}_{\mathcal{F}}(\mathcal{D}_i, Q_i)\right],$$

and for the sub-task ERM $\widehat{\psi} \in \arg\min_{\psi\in\Psi}\sum_{i=1}^{T}\widehat{R}_{S_i}(\psi)$, with $\psi^\star \in \arg\min_{\psi\in\Psi}\sum_{i=1}^{T}R_{\mathcal{D}_i}(\psi)$,

$$R_{\mathcal{D}}(F_{\widehat{\psi}}) \le \frac{B_\ell k}{\log 2}\left[\sum_{i=1}^{T}R_{\mathcal{D}_i}(\psi^\star) + 4T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) + 6B_\ell\sqrt{\frac{2T\log(4/\delta)}{n}} + 2\sum_{i=1}^{T}\mathrm{disc}_{\mathcal{F}}(\mathcal{D}_i, Q_i)\right].$$

The proof is given in Appendix B.6. The corollary thus controls shifts between training and target step distributions; the drift of the model's own generated trajectory (its rollout) is handled only through the hard-decoding reduction and is not itself a distributional guarantee.

4 Specific Analysis: Results in the Next-Token Prediction Setting

In the preceding analysis the OOD generalization error of complex compositional tasks is upper-bounded by a sum of per-step in-distribution risks, with the gap governed by the Rademacher complexity of the shared hypothesis class. We now specialize this decomposition to the Next-Token Prediction (NTP) pretraining paradigm [23]—the standard training regime for LLMs—and derive a target-token bound with explicit depth, sample-size, and decoder-capacity dependence. Relative to the general analysis of Section 3.2 the additional ingredients are (i) a reduction from the final-task risk of free-running generation, in which the model conditions on its own outputs, to target-token risks under teacher forcing, in which every prefix is taken from the ground truth (Lemma 10, Appendix B.7) and (ii) an explicit evaluation of the complexity term for the decoder (Appendix A.2); the ERM comparison is the one already used in Corollary 3.

4.1 Problem Formulation

The model is a composition $\mathrm{LM} = g \circ h$ (representation learner h and token predictor g). Here, the abstract single-step mapping ψ is instantiated by the block predictor $\psi_{g,h}$ induced by k successive next-token predictions via $g \circ h$, and $F_{\psi_{g,h}}$ denotes the resulting T-step final predictor. For step $i = 1, \ldots, T$, let $S_i = \{X_{ij}\}_{j=1}^n$, where each $X_{ij} = (t_{ij,1}, \ldots, t_{ij,m})$ is a complete sequence. We denote the teacher-forced context from X_{ij} before token t by $X_{ij,<t}$ and set $N = nT$. Here and in Appendix A, m denotes the complete-record length: an $(m-k)$-token state followed by k target tokens. Thus the general setup of Section 2 is instantiated with state length $m - k$. Before each block, both oracle and generated histories retain their most recent $m - k$ tokens; within the block, the corresponding true or generated prefixes are appended autoregressively. The oracle record is therefore $X = z_{i-1} \oplus \Delta z_i$, and $\mathcal{D}_i$ also denotes its induced distribution. This length convention leaves the general risk bounds unchanged.

For a complete sequence $X = (t_1, \ldots, t_m)$ and its prefix $X_{<t} = (t_1, \ldots, t_{t-1})$, define the target-token loss, with g applied to the hidden state at the last position of its argument:

$$q_{\psi_{g,h}}(X) := \frac{1}{k} \sum_{t=m-k+1}^{m} \ell(g(h(X_{<t})), t_t)\,, \tag{1}$$

and let $\mathcal{Q}_k = \{q_{\psi_{g,h}} : g \in \mathcal{G},\ h \in \mathcal{H}\}$, where $\mathcal{G}, \mathcal{H}$ are the classes of the token predictor and of the representation learner. For $S_i = \{X_{ij}\}_{j=1}^n$, define

$$R^k_{\mathcal{D}_i}(\psi_{g,h}) := \mathbb{E}_{X\sim\mathcal{D}_i}[q_{\psi_{g,h}}(X)], \qquad \widehat{R}^k_{S_i}(\psi_{g,h}) := \frac{1}{n}\sum_{j=1}^{n} q_{\psi_{g,h}}(X_{ij}).$$

The $N = nT$ complete sequences are the independent sampling units; tokens and prefixes within each sequence may be dependent. We abbreviate $R^k_i := R^k_{\mathcal{D}_i}$ and $\widehat{R}^k_i := \widehat{R}^k_{S_i}$.

Remark 6 (Teacher-Forced NTP and Free-Running Generation). The target-token loss in Eq. (1) is *teacher-forced*: it uses the true within-step prefixes, whereas $F_{\psi_{g,h}}$ runs freely and conditions on its own generated tokens. These block losses need not coincide. Lemma 10 in Appendix B.7 extends the first-error argument of Theorem 1 to individual tokens and gives $R_{\mathcal{D}}(F_{\psi_{g,h}}) \le \frac{B_\ell k}{\log 2} \sum_{i=1}^{T} R^k_{\mathcal{D}_i}(\psi_{g,h})$ with the same coefficient $B_\ell k/\log 2 = \mathcal{O}(k)$.

In this setting we write $(\widehat{g}, \widehat{h})$ for the minimizer of $\sum_i \widehat{R}^k_{S_i}(\psi_{g,h})$ and $(g^\star, h^\star)$ for the minimizer of $\sum_i R^k_{\mathcal{D}_i}(\psi_{g,h})$, with $\widehat{\psi} := \psi_{\widehat{g},\widehat{h}}$ and $\psi^\star := \psi_{g^\star,h^\star}$; the resulting target-token excess is the NTP analogue of the atomic excess of Section 3.2.

4.2 Main Theoretical Results

Let $\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k)$ denote the empirical Rademacher complexity of the target-token class on the combined sample S_0 of $N = nT$ records, i.e., the quantity of Appendix B.1 with $\mathcal{F}$ replaced by $\mathcal{Q}_k$ and one

independent sign per record. Applying Theorem 2 to the target-token class yields the following deviation bound, uniform over the class.

Theorem 5 (Target-Token Risk Bound)**.** *Under Assumption 1, with probability at least* $1-\delta$*, uniformly over* $g \in \mathcal{G}$ *and* $h \in \mathcal{H}$*,*

$$\left| \sum_{i=1}^{T} \left(R^k_{\mathcal{D}_i}(\psi_{g,h}) - \widehat{R}^k_{S_i}(\psi_{g,h}) \right) \right| \le 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) + 3B_\ell \sqrt{\frac{2T\log(4/\delta)}{n}}. \tag{2}$$

The next corollary is the target-token counterpart of the ERM form of Corollary 3: for the empirical risk minimizer the empirical risks in the bound are replaced by the best-in-class population risks, at the price of paying the deviation twice.

Corollary 6 (ERM Final-Task Risk)**.** *Let* $(\widehat{g}, \widehat{h})$ *and* $(g^\star, h^\star)$ *be defined as above. With probability at least* $1-\delta$*,*

$$R_{\mathcal{D}}(F_{\psi_{\widehat{g},\widehat{h}}}) \le \frac{B_\ell k}{\log 2} \left[\sum_{i=1}^{T} R^k_{\mathcal{D}_i}(\psi_{g^\star,h^\star}) + 4T\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) + 6B_\ell \sqrt{\frac{2T\log(4/\delta)}{n}} \right]. \tag{3}$$

Under Assumption 2, the target-token complexity obeys the following bound. Deriving it (Appendix A.2) requires two ingredients beyond the general analysis of Section 3.2: the layerwise matrix-covering argument of [23], which we instantiate for our decoder block and then restrict to the final k output rows, and a sequence-level loss-covering step of our own that charges one Rademacher sign per complete sequence rather than per generated token; the latter is what produces the $1/\sqrt{N}$ normalization with $N = nT$.

$$\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) \lesssim G_\ell \sqrt{\frac{\Theta d H \tau_{m,N}}{N}} + G_\ell \sqrt{\frac{d n_v}{N}}, \qquad N = nT, \tag{4}$$

where Θ counts trainable parameters, L, H, d are the layer count, head count and model dimension, and $\tau_{m,N} = \log(1 + \rho_L s_L(m,N))$ is the decoder-capacity factor (Appendix A.2); below we write τ for $\tau_{m,N}$.

The resulting final-task bound is, for constants $C_1, C_2 > 0$ depending only on the fixed norm and normalization bounds of Assumption 2,

$$R_{\mathcal{D}}(F_{\psi_{\widehat{g},\widehat{h}}}) \le \frac{B_\ell k}{\log 2} \Big[\sum_{i=1}^{T} R^k_i(\psi^\star) + C_1 G_\ell \sqrt{\frac{T\Theta d H \tau}{n}} + C_2 G_\ell \sqrt{\frac{T d n_v}{n}} + 6B_\ell \sqrt{\frac{2T\log(4/\delta)}{n}} \Big]. \tag{5}$$

Since $B_\ell k / \log 2 = \mathcal{O}(k)$, for fixed model dimensions, norm bounds and clipping threshold the sample-dependent atomic excess contributes $\widetilde{\mathcal{O}}(k\sqrt{T/n})$ to the bound, where $\widetilde{\mathcal{O}}$ suppresses the logarithmic factors of τ and of the confidence term.

4.3 Discussion of the Final Bound

Depth and sample size. The target-token analysis treats the $N = nT$ complete sequences as the independent observations while allowing arbitrary dependence among tokens within each sequence. Pooling these observations through one shared predictor under a uniform capacity control on the class yields an atomic excess of order $\sqrt{T/n} = \sqrt{T} \cdot n^{-1/2}$—a $\sqrt{T}$ accumulation across the T composed steps, sublinear in place of the naive linear T, times the $n^{-1/2}$ statistical cost of a single atomic step—rather than the $T/\sqrt{n}$ dependence obtained by analyzing the T steps separately. We state this with the per-step sample size n held fixed, so that the depth dependence is $\sqrt{T}$; if the total budget $N = nT$ is held fixed instead, then n shrinks and the per-step sampling factor alone grows like $\sqrt{T/N}$, giving $T/\sqrt{N}$—linear behavior that comes from the shrinking per-step budget, not from any loss of the pooling gain. Equation (3) retains the best-in-class step-level risk in the complete final-task risk bound; after the depth-independent factor $B_\ell k / \log 2 = \mathcal{O}(k)$, the sample-dependent atomic excess contributes $\widetilde{\mathcal{O}}(k\sqrt{T/n})$ to the bound.

Context length and decoder capacity. We analyze the final-k-token loss directly, which permits dependence among tokens within a sequence; concentration is taken over the $N = nT$ independent complete sequences. In the parameterized bound of Eq. (5), context length enters through τ evaluated at $N = nT$ and also through Θ if the trainable positional-embedding parameter count grows with m.

Theoretical context Unlike prior work that restricts compositional analysis to 2-step cases without explicit parameter-dependent bounds [25], focuses on sub-task identification sample complexity rather than OOD error decomposition [26], proves that CoT reduces learning difficulty without generalization bounds [16], or addresses only single-distribution pretraining [23], our framework unites CoT error decomposition with a shared-hypothesis structure and directly controls target-token risk with reasoning-depth and decoder-capacity dependence. Maurer et al. [24] provide a shared-representation perspective for pooled estimation, whereas Li et al. [23] provide decoder covering-number tools for full-sequence NTP. We use the latter only to control the restricted final-k output class; the target-token risk bound itself is derived directly at the sequence level.

5 Experiments

We evaluate compositional generalization on Atom-Task, a task of multi-step operations on sequences; GSM8K experiments, extended setup details, and noise-robustness results are in Appendix C (GSM8K results in Section C.2).

Experimental Setup We introduce Atom-Task to isolate compositional reasoning, requiring sequential execution of atomic operations on symbolic sequences over a six-symbol alphabet ($\{\alpha, \beta, \gamma, \delta, \varepsilon, \omega\}$). We define five operations: Reverse, Substitute, Duplicate, Deduplicate, and Rotate.

Models and Datasets For the Atom-Task, we use Qwen3-4B. We design four training datasets: (1) Single-Step: 5k uniformly distributed one-step samples; (2) Three-Step: 1.66k three-step samples; (3) Mixed 3+5: 1,333 samples (833 three-step and 500 five-step; the latter use ≤ 2 distinct operations, bridging length and combinatorial generalization); (4) Mixed 1-, 3-, and 5-Step: 4,000 one-step, 600 three-step, and 400 five-step samples (for each operation *op*, a dedicated 5k-sample dataset excludes *op* from all multi-step chains but keeps it in single-step data). For dataset-size accounting, an L-step trajectory contributes L step-level samples; this convention is used only to control the training-data budget. All models are evaluated on held-out 3-, 5-, 7-, and 9-step test sets (500 samples each). Full implementation details are given in Appendix C.

Table 1: Accuracy (%) by training data and test steps. All entries are means over 3 seeds (42, 2026, 12345); standard deviations are shown as subscripts.

(a) CoT Setting

Data	3-step	5-step	7-step	9-step
1-step	$91.80_{\pm 2.24}$	$63.13_{\pm 4.15}$	$36.27_{\pm 7.87}$	$22.27_{\pm 3.67}$
3-step	$99.33_{\pm 0.25}$	$95.73_{\pm 0.68}$	$88.87_{\pm 1.54}$	$79.27_{\pm 3.03}$
Mix_{3+5}	$100.00_{\pm 0.00}$	$97.87_{\pm 0.68}$	$94.00_{\pm 0.59}$	$90.40_{\pm 1.57}$

(b) No-CoT Setting

Data	3-step	5-step	7-step	9-step
1-step	$10.93_{\pm 0.84}$	$7.00_{\pm 0.00}$	$2.73_{\pm 0.25}$	$0.67_{\pm 0.34}$
3-step	$75.87_{\pm 0.84}$	$30.20_{\pm 0.65}$	$11.00_{\pm 0.85}$	$5.13_{\pm 0.09}$
Mix_{3+5}	$51.00_{\pm 0.75}$	$26.60_{\pm 1.40}$	$12.20_{\pm 0.75}$	$5.73_{\pm 0.09}$

5.1 Main Results and Analysis

Short-to-Long Generalization & CoT vs. No-CoT. Table 1 compares CoT and No-CoT settings. While CoT with single-step data shows some compositionality, its degradation at longer test steps indicates the benefit of short-chain training data. Crucially, models trained on short chains (3-step/3+5 mixed) perform well on longer tasks; since the 5-step chains limit operation types, this implies models generalize stepwise decomposition patterns. In contrast, No-CoT degradation indicates a failure to capture compositional structures without intermediate steps, causing overfitting to spurious correlations.

Generalization to Unseen Combinations. Table 2(a) shows performance when specific operations are excluded from multi-step samples. Even with 9-step test sequences, all experimental groups maintain high accuracy, indicating that models can effectively compose and generalize operations encountered only in single-step data to multi-step tasks, thereby achieving compositional generalization. Table 2(b) again confirms the necessity of CoT: without it, performance remains consistently poor regardless of the excluded operation.

Table 2: Atom-Task: Accuracy (%) when an operation is excluded from multi-step samples. Entries are means over 3 seeds (42, 2026, 12345); standard deviations are shown as subscripts. Rev: Reverse, Dedup: Deduplicate, Sub: Substitute, Dup: Duplicate, Rot: Rotate.

(a) CoT Setting

Op	3-step	5-step	7-step	9-step
Rev	$99.53_{\pm 0.34}$	$97.13_{\pm 0.25}$	$92.67_{\pm 1.60}$	$87.13_{\pm 0.84}$
Dedup	$97.87_{\pm 0.47}$	$93.27_{\pm 1.11}$	$86.20_{\pm 1.82}$	$76.73_{\pm 0.66}$
Sub	$96.80_{\pm 0.43}$	$90.73_{\pm 0.41}$	$84.33_{\pm 0.38}$	$70.67_{\pm 0.81}$
Dup	$93.60_{\pm 0.59}$	$84.80_{\pm 0.99}$	$76.73_{\pm 1.75}$	$66.73_{\pm 1.25}$
Rot	$96.53_{\pm 0.66}$	$91.93_{\pm 0.66}$	$86.93_{\pm 0.34}$	$80.13_{\pm 0.41}$

(b) No-CoT Setting

Op	3-step	5-step	7-step	9-step
Rev	$53.73_{\pm 0.50}$	$30.33_{\pm 0.74}$	$16.00_{\pm 1.18}$	$5.53_{\pm 0.47}$
Dedup	$40.40_{\pm 0.65}$	$16.93_{\pm 0.66}$	$6.80_{\pm 0.43}$	$1.53_{\pm 0.09}$
Sub	$42.87_{\pm 0.81}$	$21.40_{\pm 0.65}$	$11.20_{\pm 0.33}$	$5.07_{\pm 0.50}$
Dup	$38.87_{\pm 1.23}$	$17.20_{\pm 0.75}$	$11.07_{\pm 1.09}$	$3.40_{\pm 0.57}$
Rot	$55.47_{\pm 0.41}$	$35.60_{\pm 1.84}$	$22.67_{\pm 1.09}$	$9.73_{\pm 0.93}$

CoT with Full State

```
Step 1: Task_Chain: [Dup, Rev, Sub] | Input:
[xxx] | → Output: [xxx] | Task_Chain: [Rev, Sub]
Step 2: Task_Chain: [Rev, Sub] | Input:
[xxx] | → Output: [xxx] | Task_Chain: [Sub]
Step 3: Task_Chain: [Sub] | Input:
[xxx] | → Output: [xxx] | Task_Chain: []
Final Answer: [xxx]
```

CoT without Full State

```
Step 1: [Dup] Task_Chain: [..., ..., ...] | Input:
[...] | → Output: [xxx] | Task_Chain: [..., ...]
Step 2: [Rev] Task_Chain: [..., ...] | Input:
[...] | → Output: [xxx] | Task_Chain: [...]
Step 3: [Sub] Task_Chain: [...] | Input:
[...] | → Output: [xxx] | Task_Chain: []
Final Answer: [xxx]
```

Figure 2: Comparison of complete vs. incomplete CoT state recordings.

Effect of State Tracking in CoT. We ablate the CoT format by removing explicit state tracking (remaining tasks and input context) from the intermediate outputs (shown in Figure 2), such that the intermediate state z_{i-1} can no longer carry complete information. Table 3 shows this ablation causes a consistent performance degradation (see the drop in average accuracy on the 7-step test set). This demonstrates that when z_{i-1} fails to preserve the full state, multi-step decomposition is impaired. This validates our theoretical requirement that z_{i-1} must carry complete state information.

Table 3: Format ablation: Accuracy (%) with vs. without explicit state tracking. Abl.: means over 3 seeds (42, 2026, 12345); Orig.: means over 3 seeds from Table 2.

Step	Reverse	Rotate	Duplicate	Deduplicate	Substitute	Avg. (Abl.)	Avg. (Orig.)
3-step	$98.47_{\pm 1.05}$	$95.53_{\pm 2.08}$	$92.33_{\pm 4.97}$	$96.93_{\pm 1.43}$	$94.53_{\pm 4.20}$	$95.56_{\pm 1.41}$	$96.87_{\pm 0.23}$
5-step	$95.80_{\pm 2.69}$	$88.13_{\pm 5.12}$	$83.20_{\pm 6.95}$	$91.73_{\pm 5.49}$	$87.40_{\pm 7.08}$	$89.25_{\pm 2.55}$	$91.57_{\pm 0.34}$
7-step	$79.40_{\pm 10.16}$	$67.80_{\pm 2.61}$	$67.47_{\pm 6.13}$	$62.47_{\pm 4.43}$	$60.67_{\pm 20.02}$	$67.56_{\pm 4.77}$	$85.37_{\pm 0.61}$

Scaling Trends of Compositional Generalization We study performance as a function of reasoning depth, training sample size, and the number of inference-time examples, as summarized in Figure 3. For depth scaling (Figure 3a), error increases with T, with diminishing increments for single-step training. Theorem 2 gives sublinear-in-T control of the estimation term at fixed n. For sample scaling (Figure 3b), increasing the sample size consistently reduces the test error rate, particularly for longer reasoning chains, which is qualitatively consistent with a smaller estimation error at larger sample sizes. Figure 3c varies the number of in-context examples at inference time rather than the theoretical sliding-window length m. Detailed data for these scaling experiments are provided in Appendix C.1.3.

6 Further Related Work

CoT Prompting and Training Chain-of-Thought (CoT) enhances complex problem-solving by generating intermediate reasoning steps [1], improving performance on math benchmarks [27]. This led to extensions such as Tree of Thoughts [28], Graph of Thoughts [29], and Program of Thoughts [30], alongside robustness improvements via self-consistency [31], self-refinement [32], and retrieval-augmented generation [33]. Decomposition methods like Least-to-Most prompting [34] and Automatic CoT [35] further automate this process. However, most work focuses on inference-stage prompts, with limited discussion of CoT's intrinsic training mechanisms [18]. Training LLMs with CoT via supervised fine-tuning or reinforcement learning, on the other hand, enhances capability

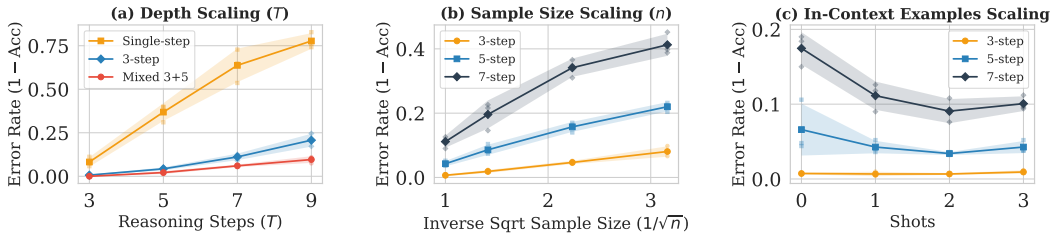


Figure 3: Each entry is a mean over 3 seeds (42, 2026, 12345). Atom-Task results when varying (a) test depth, (b) training sample size, and (c) inference-time examples; in (b) the horizontal axis is $1/\sqrt{n}$, where n is the sample size in units of the full budget (10%, 20%, 50%, 100%).

[3, 5, 36] and improves compositional generalization [25]; a reasoning trace also leaves a transferable representation-level signature [37].

CoT Theoretical Analysis Theoretically, CoT increases Transformers' effective depth for serial problems [14, 15], and linear next-token predictors on CoT data can approximate Turing-computable functions [13]. Further analyses cover sample/time complexity [38], task decomposition [16], in-context learning bounds [17], and expressive power on arithmetic or gradient descent [39], with structural analysis showing reasoning internalized as combinatorial circuits [40, 41]. However, these studies are often qualitative or limited to localized tasks [11, 18]. More specifically, Joshi et al. [38] and Hanneke et al. [42] analyze the PAC (probably approximately correct) error of the final symbol under complete-trajectory supervision and realizability, whereas our setting supervises individual atomic steps through a target-token objective and provides explicit depth-dependent bounds. The analysis of [25] considers out-of-distribution generalization but is restricted to two-step tasks and lacks LLM-specific analysis. Maurer et al. [24] provide a shared-representation perspective for pooled estimation, whereas Li et al. [23] provide the decoder covering-number tool used for our restricted final-k output class.

7 Conclusion and Future Work

In this work we asked where the out-of-distribution capability of CoT comes from, and located it in the reuse of one transition model along the chain. The analysis separates a multiplicative price for inference errors, of order $\mathcal{O}(k)$, set by the block length rather than by the number of steps, from a statistical cost that grows only as $\mathcal{O}(\sqrt{T/n})$, since a single predictor serves all T sub-tasks and can be assessed on a combined sample. In the token-by-token prediction regime this gives an end-to-end bound that is $B_\ell k/\log 2$ times the best-in-class summed step-level risk, plus an estimation term of order $\widetilde{\mathcal{O}}(k\sqrt{T/n})$; a matched construction for the sub-task ERM shows that the $\sqrt{T/n}$ dependence is not an artefact of the analysis. The decomposition further indicates how learning reusable transitions can support compositional transfer—consistent with our experiments, where explicit state tracking proves essential. Taken together, the analysis sharpens our picture of how compositional OOD generalization arises under CoT, and that picture is not tied to autoregressive token-by-token output: it may also inform future analyses of agents that maintain and update an external memory.

Limitations & Future Work Our theory studies explicit CoT under supervised NTP-style training, where intermediate states and step-level transitions are observable; it excludes adaptive paradigms such as reinforcement fine-tuning and test-time search [11, 43–46], where trajectories may be selected, refined, or rewarded. Extending our decomposition framework to these settings is a natural future direction, and does not affect the mechanism identified here.

Acknowledgments and Disclosure of Funding

This work was supported by the Special Fund for Key Program of Science and Technology of Jiangsu Province (No. BG2024042); the Gusu Leading Talents Program for Innovation and Entrepreneurship (No. ZXL2025323); the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the

Ministry of Education of China (No. JYB2025XDXM118); and the "111 Center" (No. B26023). This work was also partially supported by the Lee Kong Chian Fellowship awarded to Pan Zhou by Singapore Management University. We thank the NeurIPS reviewers for their valuable feedback and suggestions.

References

[1] 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.

[2] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[3] Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024.

[5] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[6] Fangxu Yu. Flow of reasoning: Efficient training of llm policy with divergent thinking. *arXiv (Cornell University)*, 2024.

[7] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882, 2023.

[8] Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, 2023.

[9] Team OpenAI. Learning to reason with llms. *OpenAI Blog*, 2024.

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. In *Second Conference on Language Modeling*, 2025.

[12] Akshara Prabhakar, Thomas L Griffiths, and R Thomas McCoy. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724, 2024.

[13] 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.

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

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

[16] Chenxiao Yang, Zhiyuan Li, and David Wipf. Chain-of-thought provably enables learning the (otherwise) unlearnable. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Alireza Amiri Bavandpour, Xinting Huang, Mark Rofin, and Michael Hahn. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *International Conference on Machine Learning*, pages 3274–3306. PMLR, 2025.

[18] 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.

[19] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] Liyuan Liu, Biqin Song, Zhibin Pan, Chuanwu Yang, Chi Xiao, and Weifu Li. Gradient learning under tilted empirical risk minimization. *Entropy*, 24(7):956, 2022.

[21] Yuyang Deng, Junyuan Hong, Jiayu Zhou, and Mehrdad Mahdavi. On the generalization ability of unsupervised pretraining. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4527. PMLR, 2024.

[22] Liyuan Liu, Hong Chen, Chi Xiao, and Weifu Li. The consistency analysis of gradient learning under independent covariate shift. *Neurocomputing*, 635:129883, 2025.

[23] Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *Forty-second International Conference on Machine Learning*, 2025.

[24] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[25] 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 *The Fourteenth International Conference on Learning Representations*, 2026.

[26] Amirhesam Abedsoltan, Huaqing Zhang, Kaiyue Wen, Hongzhou Lin, Jingzhao Zhang, and Mikhail Belkin. Task generalization with autoregressive compositional structure: Can learning from d tasks generalize to d^T tasks? *arXiv preprint arXiv:2502.08991*, 2025.

[27] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[28] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[29] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

[30] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

[31] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[34] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[37] Ziyi Wang, Li Li, Aolin Zhou, Yankun Shen, Chonghan Liu, Shuxia Lin, and Xu Yang. Can linguistic reasoning vectors enhance multimodal reasoning ability? *arXiv preprint arXiv:2609.31140*, 2026.

[38] Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 3161–3212. PMLR, 2025.

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

[40] Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems*, 37:95238–95265, 2024.

[41] Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, 2024.

[42] Steve Hanneke, Idan Mehalel, and Shay Moran. Sample complexity of autoregressive reasoning: Chain-of-thought vs. end-to-end. *arXiv preprint arXiv:2604.12013*, 2026.

[43] Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*, 2025.

[44] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or rl is suboptimal. In *International Conference on Machine Learning*, pages 54058–54094. PMLR, 2025.

[45] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[46] Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, et al. Rethinking reflection in pre-training. *arXiv preprint arXiv:2504.04022*, 2025.

A Supplementary Details on NTP Setting

This section supplements Section 4 by defining the decoder architecture and the sampling units used in the target-token analysis. We use the convention of Section 4.1: m denotes the complete-record length (and the maximum sequence length used by the decoder), the block-boundary state has length $m-k$, and $\mathcal{V}$ denotes the vocabulary.

A.1 Decoder-only Model Architecture

Contemporary Large Language Models (LLMs) predominantly utilize a decoder-only Transformer architecture. As described in Section 4.1, the composite model $\mathrm{LM} = g \circ h$ consists of a representation learner h and a token predictor g. We formalize the representation learner h as an L-layer Transformer decoder, following the block description of [23, Eq. (4)], mapping an input token sequence $Z \in \mathcal{V}^m$ to a hidden representation $h(Z) \in \mathbb{R}^{m \times d}$. The decoder applies to any prefix $Z \in \mathcal{V}^s$ with $s \le m$, and the prediction head reads its last hidden state.

The process begins with an Embedding Layer, which converts discrete tokens into continuous vectors $Z_0 \in \mathbb{R}^{m \times d}$ via:

$$Z_0 = \text{Embedding}(Z) := \text{OneHot}(Z) W_e + W_p,$$

where $\text{OneHot}(Z) \in \{0,1\}^{m \times n_v}$ is the one-hot encoding of the m tokens, $W_e \in \mathbb{R}^{n_v \times d}$ is the token embedding matrix, and $W_p \in \mathbb{R}^{m \times d}$ is the position embedding matrix. The core of the architecture consists of L stacked Transformer-decoder blocks. For layer $l \in \{1, \dots, L\}$, the output Z_l is computed from Z_{l-1} via masked self-attention and feed-forward networks (FFN). Specifically, let H be the number of attention heads and $d_k = d/H$. The masked self-attention head $h \in [H]$ is defined as:

$$\text{Attn}_h(Z_{l-1}) = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}} + M\right) V_h, \quad Q_h = Z_{l-1} W_h^Q, K_h = Z_{l-1} W_h^K, V_h = Z_{l-1} W_h^V.$$

Here, $M \in \mathbb{R}^{m \times m}$ is a causal mask matrix where $M_{ij} = -\infty$ if $j > i$ and 0 otherwise. This mask ensures that the hidden state at position t attends only to tokens at positions $\le t$, so that the token at position t is predicted from the state at position $t-1$, thereby implementing the autoregressive property essential for Next-Token Prediction (NTP). The final output of the layer is given by:

$$Y_l := \text{Norm}_1\left(\sum_{h=1}^{H} \text{Attn}_h(Z_{l-1}) W_h^O + Z_{l-1}\right), \quad Z_l = \text{Norm}_2\left(\text{FFN}(Y_l) + Y_l\right),$$

where $\text{Norm}_1, \text{Norm}_2$ denote Layer Normalization and $\text{FFN}(Y) = \sigma(Y W_{F1}) W_{F2}$ is a two-layer feed-forward network with hidden width $d_f = 4d$. The representation learner thus serializes the input into a deep representation $h(Z) = Z_L$.

The token predictor g applies a linear projection to the hidden state at the last position, $g(h(Z)) = \text{softmax}((h(Z))_{s,:} W_P)$ with $s = |Z|$ and $W_P \in \mathbb{R}^{d \times n_v}$, mapping it to a probability distribution over the vocabulary $\mathcal{V}$.

A.2 Explicit Parameterization and Derivation of the Dimension-Dependent Bound

To quantify the Rademacher term in Corollary 6, we use the matrix-covering tools recorded in [23, Lemma C.3]. The statistical argument below is carried out directly for sequence-level losses.

Assumption 2 (Capacity and Regularity Constraints)**.** For the L-layer, H-head decoder-only Transformer defined in Appendix A.1 (following [23, Eq. (4) and Assumption 4.14]), we assume:

- The Layer Normalization operators Π_{norm} (the operators Norm_1 and Norm_2 of Appendix A.1) are G_π-Lipschitz continuous with respect to the 2-norm.
- For all layers $l \in [L]$ and attention heads $h \in [H]$, the scaled attention scores satisfy $\|Q_l^h (K_l^h)^\top / \sqrt{d_k}\|_{\ell_\infty} \le C_l$.
- Each attention and feed-forward parameter matrix in layer l has Frobenius norm at most B_l.

The embedding matrices and output head also have bounded Frobenius norms. Let $B_* \geq 1$ be a common upper bound for these norms and the B_l. For the capacity estimate, normalization is a fixed rowwise operator whose output rows have norm at most $C_\pi\sqrt{d}$, with fixed C_π; the feed-forward activation is 1-Lipschitz and vanishes at zero, as does ReLU. These are the usual bounded-row normalization and feed-forward conditions. The constants B_*, G_π, C_π are fixed independently of the sample size and model dimensions. Let Θ count all trainable scalar parameters, including token and positional embeddings whenever they are trained, and the output head.

The norm and Lipschitz upper bounds may be enlarged so that $B_l, G_\pi \geq 1$. Let $N = nT$ and $\rho_L = \sum_{l=1}^{L} B_l^2$, and set

$$s_L(m,N) = \prod_{l=1}^{L} G_\pi^2 (B_l^2+1)\left(e^{C_l} B_l^2 H\sqrt{N}\, md + 1\right),$$
$$\tau_{m,N} = \log(1+\rho_L s_L(m,N)).$$

Here s_L collects the per-layer amplification factors that a parameter perturbation of size η accumulates over the L blocks, and $\tau_{m,N}$ is the logarithm of the resulting covering-number bound; the individual factors are derived in the propagation estimate below.

For fixed norm, attention-score, and normalization bounds, $\tau_{m,N} = \mathcal{O}(L\log(2+NHmd) + \log(1+L))$. Below, constants in $\lesssim$ may depend on the fixed bounds B_*, G_π, C_π, but not on $G_\ell, N, m, k, L, d, H, n_v$.

Derivation of Eq. (4). We first record a coarse output cover, including every trainable matrix. A Frobenius ball of radius B_* in $\mathbb{R}^{a\times b}$ has an η-net, i.e., a set of points whose η-neighbourhood covers the ball, of cardinality at most $(1+2B_*/\eta)^{ab}$, by vectorization and [23, Lemma C.3]. Taking the product of these nets gives at most $(1+2B_*/\eta)^{\Theta}$ parameter representatives. Rather than composing the layerwise coverings of [23, Lemma 4.15 and Proposition 4.17], we propagate the approximation errors through the actual block of Appendix A.1, which keeps our normalization placement unchanged.

For completeness, write TF_l for block l. A normalized row has norm at most $C_\pi\sqrt{d}$, and the bounded feed-forward weights and residual therefore give a block-output row norm at most a constant times $\sqrt{d}$. Embedding rows satisfy the same type of bound. The causal softmax has Euclidean Lipschitz constant at most 1, and its attention matrix has operator norm at most $\sqrt{m}$. For two such admissible inputs, if every matrix in the two blocks differs by at most η, multiplication of the query, key, value, and output matrices, followed by the residual and feed-forward operations, gives

$$\|\mathrm{TF}_l(Z) - \mathrm{TF}'_l(Z')\|_F \leq \Gamma\big(\|Z-Z'\|_F + \sqrt{m}\,\eta\big), \qquad \Gamma = cHmd^2 \geq 1,$$

where c depends only on B_*, G_π, C_π. Indeed, the query-key input perturbation is bounded by a constant times $\sqrt{md}\|Z-Z'\|_F$, its parameter perturbation by a constant times $md\eta$, and the remaining projections contribute at most a constant times $\sqrt{md}$; the displayed Γ bounds all these factors. The embedding approximation error is at most $2\sqrt{m}\,\eta$. Induction over the L blocks and inclusion of the output head thus give, for the concatenated probability outputs on N records,

$$\|A - A'\|_F \leq \sqrt{Nm}\,\Lambda\eta, \qquad \Lambda = (L+3)\Gamma^{L+1}.$$

This argument covers trainable embeddings as part of the parameter vector; fixed embeddings require no net. Covering representatives can be taken in the norm balls, and recentering in the original function class, if necessary, only changes numerical constants.

Let Π_k select the final k prediction rows, aligned with the target labels as in Eq. (1). Every such row is computed from its teacher-forced prefix. Since

$$\|\Pi_k A - \Pi_k B\|_F \leq \|A-B\|_F,$$

a full-output cover also covers the selected rows. For one record, write $q(A)$ for its target-token loss with the labels fixed. The loss Lipschitz condition and Cauchy–Schwarz give

$$|q(A)-q(B)| \leq \frac{G_\ell}{k}\sum_{r=1}^{k}\|A_{m-k+r} - B_{m-k+r}\|_2$$
$$\leq \frac{G_\ell}{\sqrt{k}}\|\Pi_k A - \Pi_k B\|_F.$$

Hence the concatenated N-dimensional loss vectors are approximated within $G_\ell\sqrt{Nm/k}\,\Lambda\eta$ (aggregating the N records in ℓ_2). Choosing $\eta = \epsilon\sqrt{k}/(G_\ell\sqrt{Nm}\Lambda)$, and setting $\epsilon = G_\ell\sqrt{N}\,u$, yields the covering-number bound

$$\log \mathcal{N}(\mathcal{Q}_k|_{S_0}, G_\ell\sqrt{N}\,u, \|\cdot\|_2) \le \Theta \log\left(1 + \frac{2B_*\Lambda\sqrt{m/k}}{u}\right),$$

where $\mathcal{N}(\cdot,\cdot,\|\cdot\|_2)$ denotes the covering number, i.e., the smallest cardinality of such a net in the ℓ_2 norm, and $\mathcal{Q}_k|_{S_0}$ is the restriction of $\mathcal{Q}_k$ to the sample S_0. Center the loss vectors at any fixed member of $\mathcal{Q}_k|_{S_0}$. This leaves their empirical Rademacher complexity unchanged. The probability simplex has diameter $\sqrt{2}$, so each centered loss vector has norm at most $G_\ell\sqrt{2N}$. Dudley's entropy integral with the sequence-level normalization $1/N$ gives

$$\begin{aligned}\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) &\lesssim G_\ell\sqrt{\frac{\Theta}{N}}\int_0^{\sqrt{2}}\sqrt{\log\left(1 + \frac{2B_*\Lambda\sqrt{m/k}}{u}\right)}\,du \\ &\lesssim G_\ell\sqrt{\frac{\Theta}{N}}\sqrt{1 + \log\left(1 + 2B_*\Lambda\sqrt{m/k}\right)}.\end{aligned}$$

The last step uses $\int_0^{\sqrt{2}}\sqrt{\log(1+a/u)}\,du \lesssim \sqrt{1+\log(1+a)}$ for $a \ge 0$. Since $k \ge 1$, the logarithmic factor is at most a constant times $1 + \log(L+3) + (L+1)\log(2+Hmd)$. On the other hand, $B_l, G_\pi \ge 1$ and $C_l \ge 0$ give $\rho_L \ge L$ and $s_L \ge [2(1+Hmd)]^L$, hence

$$\tau_{m,N} \ge \log\left(1 + L\,[2(1+Hmd)]^L\right),$$

which dominates this factor. We have therefore obtained

$$\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) \lesssim G_\ell\sqrt{\frac{\Theta\tau_{m,N}}{N}}.$$

Because $dH \ge 1$, this also proves the coarser form used in the main text,

$$\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) \lesssim G_\ell\sqrt{\frac{\Theta dH\tau_{m,N}}{N}} + G_\ell\sqrt{\frac{dn_v}{N}}.$$

Here Θ already counts the head and any trainable embeddings, so the first term alone is valid and the second is redundant; the displayed bound is a looser consequence of the preceding estimate. The proof uses one Rademacher sign per sequence and introduces one factor G_ℓ; it never replaces these signs by independent signs for individual tokens.

Finally, $N = nT$ gives

$$T\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) \lesssim G_\ell\sqrt{\frac{\Theta dH\tau_{m,nT}T}{n}} + G_\ell\sqrt{\frac{Tdn_v}{n}}.$$

Substituting this in Eq. (3) gives Eq. (5). The coefficients $4T$ and $6B_\ell$ in the ERM bound are unchanged, and the reduction factor $B_\ell k/\log 2$ is applied once. Context length enters through $\tau_{m,nT}$ and also through Θ when the number of trainable positional-embedding parameters grows with m. □

A.3 Sampling Units

For each step i, the dataset S_i contains n independently sampled complete sequences from the corresponding step-level skill distribution $\mathcal{D}_i$. These sequences need not be extracted from a shared trajectory; each is drawn independently from some real-world instance of the same skill. Dependence among tokens and prefixes within a complete sequence is unrestricted. This independent-record view is the idealization underlying the bounds. The Atom-Task experiments use the one-step and short-chain datasets specified in Section 5, including three- and five-step chains; the L-step accounting serves only as a budget convention, not an independence assumption for transitions within a chain.

Table 4: Summary of Main Notations

Symbol	Meaning
$\mathcal{V}$	Vocabulary of size n_v.
$\mathrm{hard}(\cdot)$	Operator converting probability distributions to discrete tokens via argmax.
$\mathcal{D}, \mathcal{D}_i$	Target problem distribution, and data distribution for the i-th reasoning step.
Ψ, ψ, ψ_*	Hypothesis class for single-step functions, learned single-step function, and ground-truth mapping (not necessarily within Ψ).
$\bar{\psi}$	State transition mapping, $\bar{\psi}(z) := z[k+1:m] \oplus \mathrm{hard}(\psi(z))$, representing the input state for the next step.
$\mathcal{G}, \mathcal{H}, g, h$	Hypothesis classes and functions for the token predictor and representation learner.
$\psi_{g,h}, F_{\psi_{g,h}}$	Block predictor of k successive next-token predictions by $g \circ h$, and the resulting final T-step predictor.
$\ell, \mathfrak{L}$	Token-level loss (ϵ_ℓ-clipped cross-entropy of Assumption 1) and sequence-level loss (average over the k generated tokens).
$\epsilon_\ell, G_\ell, B_\ell$	Clipping threshold of ℓ ($\epsilon_\ell \in (0, 1/4)$), Lipschitz constant of ℓ w.r.t. the 2-norm, and bound of ℓ (hence of the sequence-level loss $\mathfrak{L}$).
$\mathcal{F}$	Sequence-loss class $\mathcal{F} = \mathfrak{L} \circ \Psi = \{f_\psi : \psi \in \Psi\}$ with $f_\psi(z, \Delta z) = \mathfrak{L}(\psi(z), \Delta z)$.
T, m, k, n	Total number of reasoning steps, state length in Section 2 (complete-record length in Section 4), number of generated tokens per step, number of samples per step.
$R_{\mathcal{D}}(F_\psi)$	Final-task population risk, $R_{\mathcal{D}}(F_\psi) = \int \mathfrak{L}(F_\psi(z_0), \Delta z_T)\, dP(z_0, \Delta z_T)$.
$\mathcal{D}', P$	$\mathcal{D}$ is the law of input–answer pairs $(z_0, \Delta z_T)$ with measure P, and $\mathcal{D}'$ the joint law of the reasoning trajectory $(z_0, \Delta z_1, \ldots, \Delta z_T)$.
$R^k_{\mathcal{D}_i}, \widehat{R}^k_{S_i}$	Target-token population and empirical risks of the block predictor at the i-th reasoning step.
$N, q_{\psi_{g,h}}, \mathcal{Q}_k$	Number of complete sequence samples, target-token loss, and its function class.
$Q_i, \mathrm{disc}_{\mathcal{F}}$	Atomic training distribution and train–target discrepancy.
$\mathcal{E}_k(\widehat{\psi})$	Target-token atomic excess risk of the learned block predictor, where $\widehat{\psi} := \psi_{\widehat{g},\widehat{h}}$.
Θ, L, d, H, n_v	Number of model parameters, number of layers, model dimension, number of attention heads, vocabulary size.
$\Pi_k, \tau_{m,N}$	Final-k row selector and decoder-capacity factor.

B Technical appendices and supplementary material

B.1 Technical Preliminaries

We provide the formal definitions for the theoretical frameworks used in our analysis.

PAC Learning We conduct analysis within the Probably Approximately Correct (PAC) framework, deriving explicit upper bounds for the generalization error that hold with probability at least $1 - \delta$.

Rademacher Complexity Rademacher complexity measures the richness of a function class $\mathcal{F}$. For a fixed sample $S = \{x_1, \ldots, x_N\}$, the empirical Rademacher complexity and expected Rademacher complexity are defined as

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(x_i)\right], \quad \mathfrak{R}_N(\mathcal{F}) = \mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{F})],$$

where σ_i are independent Rademacher random variables taking values ± 1 with equal probability.

Total Variation Distance For two probability distributions P and Q over the same measurable space $\mathcal{X}$, the total variation distance is defined as

$$\|P - Q\|_{\mathrm{TV}} = \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|.$$

Since $|\mathfrak{L}| \leq B_\ell$, this distance bounds the discrepancy of Corollary 4: $\mathrm{disc}_{\mathcal{F}}(\mathcal{D}_i, Q_i) \leq 2B_\ell \|\mathcal{D}_i - Q_i\|_{\mathrm{TV}}$.

B.2 Local Error Floor of a Hard-Decoding Error

To derive the generalization error decomposition, we first record the loss incurred by a hard-decoding error at a single step.

Lemma 7 (Local Error Floor of a Hard-Decoding Error)**.** *For the autoregressive predictor ψ, every state s and every ground-truth block Δs satisfy*

$$\mathfrak{L}(\psi(s), \Delta s) \;\geq\; \frac{\log 2}{k}\, \mathbb{1}\{\mathrm{hard}(\psi(s)) \neq \Delta s\},$$

where $(\log 2)/k$ is a lower bound on the sequence-level loss $\mathfrak{L}$ whenever the hard decoding differs from the ground truth.

Proof. Fix a state s and a ground-truth block Δs, and consider a token whose ground-truth index is p and whose argmax index is $q \neq p$. Then $\hat{y}_q \geq \hat{y}_p$, and since probabilities are non-negative and sum to at most 1:

$$1 \geq \hat{y}_p + \hat{y}_q \geq 2\hat{y}_p \implies \hat{y}_p \leq 1/2.$$

Since $\max\{\hat{y}_p, \epsilon_\ell\} \leq 1/2$ whenever $\hat{y}_p \leq 1/2$ and $\epsilon_\ell < 1/4$, the token-level clipped cross-entropy of every mispredicted token satisfies $\ell = -\log \max\{\hat{y}_p, \epsilon_\ell\} \geq \log 2$. A block is hard-decoding-erroneous only if at least one of its k tokens is mispredicted, so averaging over the k tokens yields $\mathfrak{L}(\psi(s), \Delta s) \geq \frac{\log 2}{k} > 0$, which is the claim. □

Remark 7. The bound uses deterministic hard predictions: if the hard predictions at a step match the ground truth then no error is recorded, and otherwise the deviation is charged through the lower bound $(\log 2)/k$ on the local loss. This is the fact used by the risk decomposition of Theorem 1.

B.3 Proof of Theorem 1 (Generalization Error Decomposition)

Proof. We aim to bound the generalization error $R_{\mathcal{D}}(F_\psi)$. Let $\mathbf{z} = (z_0, \Delta z_1, \ldots, \Delta z_T) \sim \mathcal{D}'$ and let $p(\mathbf{z})$ denote its density. Recall that the oracle state is written z_i, whereas the free-running (FR) state generated by the model is written with a hat, $\widehat{z}_0 = z_0$ and $\widehat{z}_i = \bar{\psi}(\widehat{z}_{i-1})$, so that

$$R_{\mathcal{D}}(F_\psi) = \int \mathfrak{L}\big(\psi(\widehat{z}_{T-1}), \Delta z_T\big)\, dp(\mathbf{z}).$$

We first record a fact that will be used repeatedly: *the FR and oracle trajectories coincide exactly up to the first hard-decoding error.* Formally, let

$$u_i := \mathbb{1}\big\{\mathrm{hard}(\psi(\widehat{z}_{i-1})) \neq \Delta z_i\big\}, \qquad I := \inf\{i \geq 1 : u_i = 1\},$$

with $I = \infty$ if no error occurs. Then on the event $\{u_1 = \cdots = u_{i-1} = 0\}$ the two states agree, $\widehat{z}_{i-1} = z_{i-1}$. Indeed, $\widehat{z}_0 = z_0$, and if $\widehat{z}_{j-1} = z_{j-1}$ while $u_j = 0$, then

$$\widehat{\Delta z}_j = \mathrm{hard}(\psi(\widehat{z}_{j-1})) = \mathrm{hard}(\psi(z_{j-1})) = \Delta z_j,$$

so that

$$\widehat{z}_j = \widehat{z}_{j-1}[k{+}1{:}m] \oplus \widehat{\Delta z}_j = z_{j-1}[k{+}1{:}m] \oplus \Delta z_j = z_j.$$

The claim follows by induction. We stress that this is a consequence of hard decoding alone—before the first error the two trajectories are *identical* rather than merely close—and that no Lipschitz or stability property of ψ is used.

Next, we split the final-step risk at the first error. On $\{I > T-1\}$ the observation above gives $\widehat{z}_{T-1} = z_{T-1}$, so the integrand is exactly the step-T oracle loss; on $\{I \leq T-1\}$ it is bounded by B_ℓ. Pointwise,

$$\mathfrak{L}\big(\psi(\widehat{z}_{T-1}), \Delta z_T\big) \leq \mathfrak{L}\big(\psi(z_{T-1}), \Delta z_T\big) + B_\ell\, \mathbb{1}\{I \leq T-1\},$$

and integrating against $p(\mathbf{z})$ yields

$$R_{\mathcal{D}}(F_\psi) \leq R_{\mathcal{D}_T}(\psi) + B_\ell\, \Pr(I \leq T-1) = R_{\mathcal{D}_T}(\psi) + B_\ell \sum_{i=1}^{T-1} \Pr(I = i),$$

where the last step uses that the events $\{I = i\}$ are disjoint.

It remains to charge each first-error probability to the corresponding atomic risk. Fix $i \leq T-1$. Since $\{I = i\} = \{I \geq i\} \cap \{u_i = 1\}$, and since on $\{I \geq i\}$ the observation above gives $\widehat{z}_{i-1} = z_{i-1}$, the event $\{I = i\}$ coincides with

$$\{I \geq i\} \cap \big\{\mathrm{hard}(\psi(z_{i-1})) \neq \Delta z_i\big\}.$$

Dropping the event $\{I \geq i\}$, which can only increase the probability, and using that $(z_{i-1}, \Delta z_i)$ has law $\mathcal{D}_i$, we obtain

$$\Pr(I = i) \leq \Pr_{\mathcal{D}_i}\big(\mathrm{hard}(\psi(z_{i-1})) \neq \Delta z_i\big).$$

Finally, applying the error floor of Lemma 7 pointwise and taking the expectation over $\mathcal{D}_i$ gives

$$\Pr_{\mathcal{D}_i}\big(\mathrm{hard}(\psi(z_{i-1})) \neq \Delta z_i\big) \leq \frac{k}{\log 2}\,\mathbb{E}_{\mathcal{D}_i}\big[\mathfrak{L}(\psi(z_{i-1}), \Delta z_i)\big] = \frac{k\,R_{\mathcal{D}_i}(\psi)}{\log 2}.$$

Summing over $i = 1, \ldots, T-1$ thus gives

$$\Pr(I \leq T-1) \leq \frac{k}{\log 2}\sum_{i=1}^{T-1} R_{\mathcal{D}_i}(\psi).$$

Finally, we combine the two estimates. Since $\ell(p, y) \leq \log(1/\epsilon_\ell)$ for all p, y, we may take $B_\ell = \log(1/\epsilon_\ell)$; then $\epsilon_\ell < 1/4$ and $k \geq 1$ imply

$$\frac{B_\ell k}{\log 2} \geq 2 > 1,$$

the single term $R_{\mathcal{D}_T}(\psi)$ is dominated by $\frac{B_\ell k}{\log 2} R_{\mathcal{D}_T}(\psi)$, and hence

$$R_{\mathcal{D}}(F_\psi) \leq \frac{B_\ell k}{\log 2}\sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi).$$

Because each intermediate state z_{i-1} is deterministically constructed from the prefix $(z_0, \Delta z_1, \ldots, \Delta z_{i-1})$, the joint law of $(z_{i-1}, \Delta z_i)$ induced by $\mathcal{D}'$ is exactly the sub-task distribution $\mathcal{D}_i$; marginalizing out the remaining variables, the bound becomes

$$R_{\mathcal{D}}(F_\psi) \leq \frac{B_\ell k}{\log 2}\sum_{i=1}^{T}\int \mathfrak{L}\big(\psi(z_{i-1}), \Delta z_i\big)\, d\mathcal{D}_i(z_{i-1}, \Delta z_i),$$

which is the desired decomposition. The coefficient $\frac{B_\ell k}{\log 2}$ is thus depth-independent, with $(\log 2)/k$ the local error floor of Lemma 7. □

B.4 Proof of Theorem 2 (Rademacher Complexity Bound)

We first present the multi-distribution symmetrization lemma used in Section 3.2.

Lemma 8. *Define the population risk $R_{\mathcal{D}_i}(\psi) := \int \mathfrak{L}(\psi(z_{i-1}), \Delta z_i) d\mathcal{D}_i(z_{i-1}, \Delta z_i)$ and empirical risk $\widehat{R}_{S_i}(\psi) := \frac{1}{n}\sum_{(z_{i-1}, \Delta z_i) \in S_i} \mathfrak{L}(\psi(z_{i-1}), \Delta z_i)$. We have:*

$$\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi \in \Psi}\left(\sum\nolimits_{i=1}^{T}(R_{\mathcal{D}_i}(\psi) - \widehat{R}_{S_i}(\psi))\right)\right] \leq 2T\,\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T}[\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})],$$

where $\mathcal{F} = \mathfrak{L} \circ \Psi$, $S_0 = \bigcup_{i=1}^T S_i$, and $\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})$ is the empirical Rademacher complexity of $\mathcal{F}$ on S_0.

We now prove Lemma 8.

Proof. Let $S_1', S_2', \ldots, S_T'$ be another set of i.i.d. samples, where $S_i' \sim \mathcal{D}_i^n$ and $S_i' = \{(x_{i1}', y_{i1}'), (x_{i2}', y_{i2}'), \ldots, (x_{in}', y_{in}')\}$ for $i = 1, 2, \ldots, T$. Clearly, for all $\psi \in \Psi$, the population risk is the expectation of the empirical risk over the ghost sample:

$$R_{\mathcal{D}_i}(\psi) = \mathbb{E}_{\{S_j' \sim \mathcal{D}_j^n\}_{j=1}^T}[\hat{R}_{S_i'}(\psi)].$$

Therefore, for every $\psi \in \Psi$, we have

$$R_{\mathcal{D}_i}(\psi) - \widehat{R}_{S_i}(\psi) = \mathbb{E}_{\{S_j' \sim \mathcal{D}_j^n\}_{j=1}^T}[\hat{R}_{S_i'}(\psi) - \widehat{R}_{S_i}(\psi)].$$

Taking the supremum over $\psi \in \Psi$ on both sides, and using the fact that the supremum of an expectation is upper bounded by the expectation of the supremum (Jensen's inequality), we obtain

$$\begin{aligned}
&\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\\
=&\sup_{\psi\in\Psi}\mathbb{E}_{\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sum_{i=1}^{T}(\hat{R}_{S'_i}(\psi)-\widehat{R}_{S_i}(\psi))\right]\\
\leq&\mathbb{E}_{\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(\hat{R}_{S'_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\right].
\end{aligned}$$

Taking expectation over $\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T$ on both sides yields

$$\begin{aligned}
&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\right]\\
\leq&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(\hat{R}_{S'_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\right]\\
=&\frac{1}{n}\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}\sum_{k=1}^{n}(\mathfrak{L}(\psi(x'_{ik}),y'_{ik})-\mathfrak{L}(\psi(x_{ik}),y_{ik}))\right)\right].
\end{aligned}$$

Next, we note that for each i and k, (x_{ik},y_{ik}) and (x'_{ik},y'_{ik}) are i.i.d. variables. Therefore, we can replace them without affecting the expectation. For any specific pair (i,k), let the summation over the remaining indices be denoted by $\sum_{(i',k')\neq(i,k)}$. We have:

$$\begin{aligned}
&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left((\mathfrak{L}(\psi(x'_{ik}),y'_{ik})-\mathfrak{L}(\psi(x_{ik}),y_{ik}))\right.\right.\\
&\qquad\left.\left.+\sum_{(i',k')\neq(i,k)}(\mathfrak{L}(\psi(x'_{i'k'}),y'_{i'k'})-\mathfrak{L}(\psi(x_{i'k'}),y_{i'k'}))\right)\right]\\
=&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left((-\mathfrak{L}(\psi(x'_{ik}),y'_{ik})+\mathfrak{L}(\psi(x_{ik}),y_{ik}))\right.\right.\\
&\qquad\left.\left.+\sum_{(i',k')\neq(i,k)}(\mathfrak{L}(\psi(x'_{i'k'}),y'_{i'k'})-\mathfrak{L}(\psi(x_{i'k'}),y_{i'k'}))\right)\right].
\end{aligned}$$

Let σ_{ik} be a random variable such that $\mathbb{P}[\sigma_{ik}=1]=\mathbb{P}[\sigma_{ik}=-1]=\frac{1}{2}$. Combining the two equalities above, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left((\mathfrak{L}(\psi(x'_{ik}),y'_{ik})-\mathfrak{L}(\psi(x_{ik}),y_{ik}))\right.\right.\\
&\qquad\left.\left.+\sum_{(i',k')\neq(i,k)}(\mathfrak{L}(\psi(x'_{i'k'}),y'_{i'k'})-\mathfrak{L}(\psi(x_{i'k'}),y_{i'k'}))\right)\right]\\
=&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T,\{S'_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\mathbb{E}_{\sigma_{ik}}\sup_{\psi\in\Psi}\left(\sigma_{ik}(\mathfrak{L}(\psi(x'_{ik}),y'_{ik})-\mathfrak{L}(\psi(x_{ik}),y_{ik}))\right.\right.\\
&\qquad\left.\left.+\sum_{(i',k')\neq(i,k)}(\mathfrak{L}(\psi(x'_{i'k'}),y'_{i'k'})-\mathfrak{L}(\psi(x_{i'k'}),y_{i'k'}))\right)\right].
\end{aligned}$$

The inner expectation over σ_{ik} splits the right-hand side into the two half-terms carrying signs $+$ and $-$; the swap symmetry of the i.i.d. pair $(x_{ik},y_{ik})\leftrightarrow(x'_{ik},y'_{ik})$ used in the preceding display shows

that these two half-terms are equal, so the identity holds after averaging over σ_{ik}. Repeating this process for all $i = 1, \ldots, T$ and $k = 1, \ldots, n$, we have

$$\begin{aligned}
&\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T} \left[\sup_{\psi \in \Psi} \left(\sum_{i=1}^{T} \sum_{k=1}^{n} (\mathfrak{L}(\psi(x'_{ik}), y'_{ik}) - \mathfrak{L}(\psi(x_{ik}), y_{ik})) \right) \right] \\
=&\mathbb{E}_{\substack{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \ \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T, \\ \{\{\sigma_{ik}\}_{i=1}^T\}_{k=1}^n}} \left[\sup_{\psi \in \Psi} \left(\sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} (\mathfrak{L}(\psi(x'_{ik}), y'_{ik}) - \mathfrak{L}(\psi(x_{ik}), y_{ik})) \right) \right].
\end{aligned}$$

Finally, applying the triangle inequality to the supremum:

$$\begin{aligned}
&\sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} (\mathfrak{L}(\psi(x'_{ik}), y'_{ik}) - \mathfrak{L}(\psi(x_{ik}), y_{ik})) \\
\leq &\sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} \mathfrak{L}(\psi(x'_{ik}), y'_{ik}) \\
&+ \sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} -\sigma_{ik} \mathfrak{L}(\psi(x_{ik}), y_{ik}).
\end{aligned}$$

The joint distribution of the Rademacher signs is invariant under sign reversal. Taking expectations in the preceding inequality and using this symmetry for the second term, we get

$$\begin{aligned}
&\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{\{\sigma_{ik}\}_{i=1}^T\}_{k=1}^n} \left[\sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} (\mathfrak{L}(\psi(x'_{ik}), y'_{ik}) - \mathfrak{L}(\psi(x_{ik}), y_{ik})) \right] \\
\leq &\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{\{\sigma_{ik}\}_{i=1}^T\}_{k=1}^n} \left[\sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} \mathfrak{L}(\psi(x'_{ik}), y'_{ik}) \right] \\
&+ \mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{\{\sigma_{ik}\}_{i=1}^T\}_{k=1}^n} \left[\sup_{\psi \in \Psi} \sum_{i=1}^{T} \sum_{k=1}^{n} \sigma_{ik} \mathfrak{L}(\psi(x_{ik}), y_{ik}) \right] \\
= &2 \cdot \mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T, \{S'_j \sim \mathcal{D}_j^n\}_{j=1}^T} \left[nT \cdot \widehat{\mathfrak{R}}_{S_0}(\mathcal{F}) \right] = 2nT \cdot \mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T} [\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})].
\end{aligned}$$

Substituting this result into the previous equation involving the $1/n$ factor, we obtain the final bound:

$$\begin{aligned}
&\mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T} \left[\sup_{\psi \in \Psi} \left(\sum_{i=1}^{T} (R_{\mathcal{D}_i}(\psi) - \widehat{R}_{S_i}(\psi)) \right) \right] \\
&\leq \frac{1}{n} \cdot 2nT \, \mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T} [\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})] \\
&= 2T \, \mathbb{E}_{\{S_j \sim \mathcal{D}_j^n\}_{j=1}^T} [\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})].
\end{aligned}$$

□

Lemma 9 (McDiarmid's Inequality[19])**.** *Let V be a domain, and let $f : V^n \to \mathbb{R}$ be a function of n variables satisfying for some $c > 0$, for all $i \in [n]$ and all $x_1, \ldots, x_n, x'_i \in V$:*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c.$$

Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in V. Then with probability at least $1 - \delta$:

$$\left| f(X_1, \ldots, X_n) - \mathbb{E}\big[f(X_1, \ldots, X_n)\big] \right| \leq c \sqrt{\frac{n}{2} \ln \left(\frac{2}{\delta} \right)}.$$

Now we give the proof of Theorem 2.

Theorem 2 (Rademacher Complexity Bound). 1. Under the preceding conditions (where the step-level loss is bounded by B_ℓ), with probability at least $1-\delta$:

$$\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi)) \le 2T\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T}[\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})]+B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}.$$

2. Under the preceding conditions (where the step-level loss is bounded by B_ℓ), with probability at least $1-\delta$:

$$\left|\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right| \le 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})+3B_\ell\sqrt{\frac{2T\log\frac{4}{\delta}}{n}}.$$

Proof. We first prove the first inequality. Since $\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))$ satisfies the conditions of Lemma 9 with Tn variables and constant $2B_\ell/n$, applying Lemma 9 and then Lemma 8 yields that with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\\
\le&\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\\
\le&\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T}\left[\sup_{\psi\in\Psi}\left(\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right)\right]+B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}\\
\le&2T\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T}[\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})]+B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}.
\end{aligned}$$

For the second inequality, based on the first one, since $\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})$ also satisfies the conditions of Lemma 9 (with Tn variables, constant $2B_\ell/(Tn)$), with probability at least $1-2\delta$ we have:

$$\begin{aligned}
&\sum_{i=1}^{T}(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\\
\le&2T\mathbb{E}_{\{S_j\sim\mathcal{D}_j^n\}_{j=1}^T}[\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})]+B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}\\
\le&2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})+2T\left(\frac{2B_\ell}{nT}\sqrt{\frac{nT\log\frac{2}{\delta}}{2}}\right)+B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}\\
=&2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})+3B_\ell\sqrt{\frac{2T\log\frac{2}{\delta}}{n}}.
\end{aligned}$$

The reverse deviation follows by the same sign-symmetry argument. Using $\delta/2$ in place of δ converts the $1-2\delta$ event and the $\log(2/\delta)$ factors above into the $1-\delta$ event and the $\log(4/\delta)$ factors of part 2, without changing the coefficients. □

B.5 Proof of Corollary 3

Proof. Let $U := 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{F})+3B_\ell\sqrt{2T\log(4/\delta)/n}$, so that Theorem 2 bounds $\left|\sum_i(R_{\mathcal{D}_i}(\psi)-\widehat{R}_{S_i}(\psi))\right|$ by U uniformly over ψ, on an event of probability at least $1-\delta$.

First display. Theorem 1 gives $R_{\mathcal{D}}(F_\psi)\le\frac{B_\ell k}{\log 2}\sum_i R_{\mathcal{D}_i}(\psi)$; replacing the population risks by their empirical counterparts costs at most U, and multiplying by $B_\ell k/\log 2$ gives the first display.

Second display. Let $\widehat{\psi}$ be the sub-task ERM and $\psi^\star$ a population minimizer. Empirical optimality gives $\sum_i \widehat{R}_{S_i}(\widehat{\psi}) \le \sum_i \widehat{R}_{S_i}(\psi^\star)$, so the atomic excess of $\widehat{\psi}$ obeys

$$\sum_i R_{\mathcal{D}_i}(\widehat{\psi}) - \sum_i R_{\mathcal{D}_i}(\psi^\star) = \underbrace{\sum_i \big(R_{\mathcal{D}_i}(\widehat{\psi}) - \widehat{R}_{S_i}(\widehat{\psi})\big)}_{\le U} + \underbrace{\sum_i \big(\widehat{R}_{S_i}(\widehat{\psi}) - \widehat{R}_{S_i}(\psi^\star)\big)}_{\le 0} + \underbrace{\sum_i \big(\widehat{R}_{S_i}(\psi^\star) - R_{\mathcal{D}_i}(\psi^\star)\big)}_{\le U} \le 2U.$$

Substituting $\sum_i R_{\mathcal{D}_i}(\widehat{\psi}) \le \sum_i R_{\mathcal{D}_i}(\psi^\star) + 2U$ into the bound of Theorem 1 gives the second display. □

B.6 Proof of Corollary 4

Proof. By the oracle-prefix reduction of Theorem 1 the final-task risk is bounded by $\frac{B_\ell k}{\log 2}\sum_{i=1}^T R_{\mathcal{D}_i}(\psi)$; transferring each term via $R_{\mathcal{D}_i}(\psi) \le R_{Q_i}(\psi) + \mathrm{disc}_{\mathcal{F}}(\mathcal{D}_i, Q_i)$ and applying Theorem 2 to the Q_i proves the first display. The second follows from the same comparison with the deviation paid once for each of $\widehat{\psi}$ and $\psi^\star$, and the transfer from Q_i to $\mathcal{D}_i$ likewise paid twice. □

B.7 Teacher-Forced NTP Reduction

Lemma 10 (Teacher-Forced NTP Reduction)**.** *Under Assumption 1, consider the NTP setting of Section 4.1, with deterministic argmax decoding and the same context-window rule for teacher-forced and free-running prefixes. Then*

$$R_{\mathcal{D}}(F_{\psi_{g,h}}) \le \frac{B_\ell k}{\log 2}\sum_{i=1}^T R^k_{\mathcal{D}_i}(\psi_{g,h}).$$

The risks on the right are the teacher-forced target-token risks defined in Eq. (1), whereas the risk on the left is evaluated on the free-running final prediction.

Proof. We follow the first-error argument of Theorem 1, now locating the first erroneous token within the first erroneous block. Fix g, h, and let $\psi := \psi_{g,h}$ be the block predictor given by k successive free-running next-token predictions via $g \circ h$. As in Appendix B.3, let $\mathbf{z} = (z_0, \Delta z_1, \dots, \Delta z_T) \sim \mathcal{D}'$ with joint law $p(\mathbf{z})$, and write

$$\widehat{z}_0 = z_0, \qquad \widehat{\Delta z}_i = \mathrm{hard}(\psi(\widehat{z}_{i-1})), \qquad \widehat{z}_i = \bar{\psi}(\widehat{z}_{i-1}).$$

For $r = 1, \dots, k$, let $X_{i,<r}$ be the oracle context before token $(\Delta z_i)_r$, formed from z_{i-1} and the true within-step prefix $(\Delta z_i)_{<r}$. Let $\widehat{X}_{i,<r}$ be its free-running counterpart, formed from $\widehat{z}_{i-1}$ and $(\widehat{\Delta z}_i)_{<r}$. Both contexts use the same context-window rule, including truncation. For each oracle block, abbreviate its teacher-forced target-token loss by

$$q_i := \frac{1}{k}\sum_{r=1}^k \ell(g(h(X_{i,<r})), (\Delta z_i)_r)\,.$$

This is Eq. (1) evaluated on the oracle record for step i; consequently, $\int q_i\, dp(\mathbf{z}) = R^k_{\mathcal{D}_i}(\psi_{g,h})$.

Define, as in the proof of Theorem 1,

$$u_i := \mathbb{1}\{\widehat{\Delta z}_i \ne \Delta z_i\}, \qquad I := \inf\{i \in \{1, \dots, T\} : u_i = 1\},$$

with $I = \infty$ if no error occurs. The same induction gives $\widehat{z}_{i-1} = z_{i-1}$ on $\{I \ge i\}$. On $\{I = i\}$, let r be the first position in block i with $(\widehat{\Delta z}_i)_r \ne (\Delta z_i)_r$. Then

$$\widehat{z}_{i-1} = z_{i-1}, \qquad (\widehat{\Delta z}_i)_{<r} = (\Delta z_i)_{<r}, \qquad \widehat{X}_{i,<r} = X_{i,<r}.$$

Thus the erroneous hard prediction at this position is the argmax of the same probability vector $g(h(X_{i,<r}))$ that appears in the teacher-forced loss. The probability assigned to the true token is therefore at most $1/2$. By the clipped cross-entropy calculation in Lemma 7,

$$\ell(g(h(X_{i,<r})), (\Delta z_i)_r) \geq \log 2, \qquad q_i \geq \frac{\log 2}{k} \quad \text{on } \{I = i\}.$$

This argument includes $i = T$, so it also accounts for errors within the final output block.

Next, split the final-step loss according to whether any hard error occurs. On $\{I > T\}$, all generated tokens agree with their oracle counterparts, including every prefix inside the final block, and hence

$$\mathfrak{L}\big(\psi(\widehat{z}_{T-1}), \Delta z_T\big) = q_T.$$

On $\{I = i\}$, boundedness gives $\mathfrak{L}(\psi(\widehat{z}_{T-1}), \Delta z_T) \leq B_\ell$. Since $B_\ell k / \log 2 \geq 1$ and the events $\{I > T\}, \{I = 1\}, \ldots, \{I = T\}$ form a partition, pointwise we obtain

$$\begin{aligned}
\mathfrak{L}\big(\psi(\widehat{z}_{T-1}), \Delta z_T\big) &\leq \mathbb{1}\{I > T\}\, q_T + B_\ell \sum_{i=1}^{T} \mathbb{1}\{I = i\} \\
&\leq \frac{B_\ell k}{\log 2} \left[\mathbb{1}\{I > T\}\, q_T + \sum_{i=1}^{T} \mathbb{1}\{I = i\}\, q_i \right] \\
&\leq \frac{B_\ell k}{\log 2} \sum_{i=1}^{T} q_i.
\end{aligned}$$

Finally, integrating against $p(\mathbf{z})$ and using the oracle step marginals $\mathcal{D}_i$ yields

$$R_{\mathcal{D}}(F_{\psi_{g,h}}) \leq \frac{B_\ell k}{\log 2} \sum_{i=1}^{T} \int q_i \, dp(\mathbf{z}) = \frac{B_\ell k}{\log 2} \sum_{i=1}^{T} R^k_{\mathcal{D}_i}(\psi_{g,h}),$$

which proves the claim. The argument uses equality of the two prefixes only before the first hard error; it does not identify the teacher-forced and free-running block losses after that error. □

B.8 Proof of Theorem 5

Proof. Apply Theorem 2 to the bounded sequence-level class $\mathcal{Q}_k$ and the samples $S_i \sim \mathcal{D}_i^n$. Uniformly over $g \in \mathcal{G}$ and $h \in \mathcal{H}$,

$$\left| \sum_{i=1}^{T} \left(R^k_{\mathcal{D}_i}(\psi_{g,h}) - \widehat{R}^k_{S_i}(\psi_{g,h}) \right) \right| \leq 2T \widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) + 3B_\ell \sqrt{\frac{2T \log(4/\delta)}{n}},$$

which is Eq. (2).

For the ERM comparison, the three-term decomposition of the proof of Corollary 3 (Appendix B.5) applies verbatim with $\mathcal{F}$ replaced by $\mathcal{Q}_k$ and $R_{\mathcal{D}_i}$ by $R^k_{\mathcal{D}_i}$: writing $\mathcal{E}_k(\widehat{\psi}) := \sum_i R^k_{\mathcal{D}_i}(\widehat{\psi}) - \sum_i R^k_{\mathcal{D}_i}(\psi_{g^\star,h^\star})$ for the target-token atomic excess and $U_k := 2T\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) + 3B_\ell\sqrt{2T\log(4/\delta)/n}$, empirical optimality and the two directions of the same uniform event give $\mathcal{E}_k(\widehat{\psi}) \leq U_k + 0 + U_k$. Lemma 10 then gives

$$R_{\mathcal{D}}(F_{\psi_{\widehat{g},\widehat{h}}}) \leq \frac{B_\ell k}{\log 2} \sum_{i=1}^{T} R^k_{\mathcal{D}_i}(\widehat{\psi}) = \frac{B_\ell k}{\log 2} \left[\sum_{i=1}^{T} R^k_{\mathcal{D}_i}(\psi_{g^\star,h^\star}) + \mathcal{E}_k(\widehat{\psi}) \right],$$

which is Eq. (3) after substituting $U_k + 0 + U_k = 4T\widehat{\mathfrak{R}}_{S_0}(\mathcal{Q}_k) + 6B_\ell\sqrt{2T\log(4/\delta)/n}$. The projection and loss-cover argument in Appendix A.2 gives Eq. (4). □

B.9 Tightness of the Additive Reduction

Proposition 11 (Tightness of the Additive Reduction)**.** *For every $T \geq 1$ and every $p \in [0, 1/T]$ there is a task with the following complete description. Let $\mathcal{V} = \{0, 1\}$, $k = 1$, $m = T + 1$, and draw*

$u_1, \dots, u_T$ *i.i.d.* Bernoulli(p). *The oracle trajectory is* $z_0 = (u_1, \dots, u_T, 0)$ *with* $\Delta z_i = 0$ *for every* i*, so that the oracle state before step* i *is* $z_{i-1} = (u_i, \dots, u_T, 0, \dots, 0)$*, whose first token is* u_i *and whose last token is* 0*, and the step marginal* $\mathcal{D}_i$ *is the law of* $(z_{i-1}, 0)$ *induced by* $(u_i, \dots, u_T)$*. The shared single-step predictor is the deterministic rule* $\mathrm{hard}(\psi(z)) = \max\{(z)_1, (z)_m\}$*, with* $\psi(z)$ *putting probability* 1 *on that token, so that the clipped cross-entropy of Assumption 1 charges* 0 *when the output is correct and* $B_\ell = \log(1/\epsilon_\ell)$ *otherwise. Writing* $e_0 = 0$ *and* $e_i = e_{i-1} \vee u_i$ *for the error flag (*$\vee$ *is logical OR, and a local error puts the state in an absorbing error mode), this task satisfies*

$$R_{\mathcal{D}}(F_\psi) = B_\ell \left[1 - (1-p)^T\right] = \Omega\left(\sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi)\right), \qquad \sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi) = TpB_\ell.$$

Proof. Verify the construction of the statement. On an oracle state the last token is 0, so the hard output equals u_i and the error event is $u_i = 1$; hence $R_{\mathcal{D}_i}(\psi) = pB_\ell$. Once an error occurs, its output 1 becomes the last token of every later state and the rule keeps emitting 1, so the final output is wrong exactly when $\bigvee_{i=1}^{T} u_i = 1$, which gives

$$R_{\mathcal{D}}(F_\psi) = B_\ell \Pr\left(\bigvee_{i=1}^{T} u_i = 1\right) = B_\ell \left[1 - (1-p)^T\right].$$

The union bound gives $1 - (1-p)^T \le Tp$, and for $Tp \le 1$, $1 - (1-p)^T \ge Tp - \binom{T}{2}p^2 \ge \frac{1}{2}Tp$. Consequently

$$\frac{1}{2}TpB_\ell \le R_{\mathcal{D}}(F_\psi) \le TpB_\ell = \sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi),$$

hence $R_{\mathcal{D}}(F_\psi) = \Omega\big(\sum_{i=1}^{T} R_{\mathcal{D}_i}(\psi)\big)$, which proves the order-tightness of the additive reduction. □

What the construction shows. The witness combines two features: the local error events are independent across steps, and the state carries an absorbing error flag, so that the final output is wrong exactly when at least one local error has occurred. That event has probability $1 - (1-p)^T = \Theta(Tp)$ for $Tp \le 1$, of the same order as the sum of the T step risks, while the per-step risk itself is $R_{\mathcal{D}_i}(\psi) = pB_\ell$. The additive form of Theorem 1 is therefore tight up to its depth-independent coefficient, and the linear-in-T growth of the bound—here the genuine linear growth of the target risk—cannot be removed by any refinement of the argument.

B.10 Tightness of the Atomic Excess

Proposition 12 (Tightness of the Atomic Excess)**.** *Fix* $T = 4r + 1$ *with* $r \in \mathbb{N}$*,* $r \ge 1$*, a sign* $\theta \in \{-1, +1\}$*, and* $\gamma \in (0, 1/(2T)]$*; the sign* θ *only selects which of the two hypotheses is favoured, and the constants below do not depend on it. Let* $\mathcal{V} = \{0, 1\}$*,* $k = 1$*,* $m = T + 1$*, and let* P_θ *be the law on* $\{0, 1\}^T$ *that assigns weight* $\frac{1}{2T} + \theta\gamma$ *to each vector* $u^{(j)} \in \{0, 1\}^T$*,* $j \in [T]$*, whose support is the cyclic block* $\{j, j+1, \dots, j+2r\}$ *of* $2r + 1$ *consecutive coordinates (indices modulo* T*), and weight* $\frac{1}{2T} - \theta\gamma$ *to* $v^{(j)}$*, the same block with its leading coordinate* j *flipped, so that* $v^{(j)}$ *has ones exactly on* $\{j+1, \dots, j+2r\}$*; these* $2T$ *weights sum to one. For* $X \sim P_\theta$*,* $P_\theta(X_i = 1) = \frac{1}{2} + \theta\gamma$ *and* $P_\theta\big(\bigoplus_{i=1}^{T} X_i = 1\big) = \frac{1}{2} + \theta T\gamma$*, where* $\bigoplus$ *denotes bitwise XOR in this construction (not sequence concatenation). The oracle trajectory is* $z_0 = (X_1, \dots, X_T, 0)$ *with* $\Delta z_i = p_i$ *and* $p_i = p_{i-1} \oplus X_i$*,* $p_0 = 0$*, so that the state before step* i *exposes* X_i *as its first token and* p_{i-1} *as its last token, and* $\mathcal{D}_i^\theta$ *is the law of the* i*-th oracle state and update. The two-element class is* $\Psi = \{\psi_0, \psi_1\}$*, with* $\psi_s(z)$ *putting probability* $3/4$ *on token* $(z)_m \oplus s$ *and* $1/4$ *on the other token, so that a correct output costs* $\ell_{\mathrm{ok}} = \log(4/3)$ *and a wrong one* $\ell_{\mathrm{err}} = \log 4$ *(clipping is inactive since* $\epsilon_\ell < 1/4$*). With* n *i.i.d. draws from each* $\mathcal{D}_i^\theta$*, the sub-task ERM* $\widehat{\psi} = \arg\min_{\psi \in \Psi} \sum_{i=1}^{T} \widehat{R}_{S_i}(\psi)$ *satisfies, for* $\gamma = c_0 \min\{1/\sqrt{nT}, 1/T\}$ *with a universal constant* $c_0 \le 1/2$*,*

$$\mathbb{E}_S\left[\sum_{i=1}^{T} R^\theta_{\mathcal{D}_i}(\widehat{\psi}) - \min_{\psi \in \Psi}\Big\{\sum_{i=1}^{T} R^\theta_{\mathcal{D}_i}(\psi)\Big\}\right] = \Omega\left(\min\left\{1, \sqrt{\frac{T}{n}}\right\}\right).$$

In this construction the atomic excess coincides with the final-task excess, so the same bound holds for $\mathbb{E}_S[R^\theta_{\mathcal{D}}(F_{\widehat{\psi}}) - \min_{\psi \in \Psi} R^\theta_{\mathcal{D}}(F_\psi)]$*, and it is* $\Omega(\sqrt{T/n})$ *when* $T \le n$*.*

Proof. Verify the construction of the statement. The weights of P_θ are nonnegative and sum to one; each $u^{(j)}$ has odd parity and each $v^{(j)}$ even parity, and each coordinate lies in exactly $2r+1$ of the $u^{(j)}$ and $2r$ of the $v^{(j)}$, so that for every coordinate i

$$P_\theta(X_i = 1) = \frac{1}{2} + \theta\gamma, \qquad P_\theta\left(\bigoplus_{i=1}^{T} X_i = 1\right) = \frac{1}{2} + \theta T\gamma. \tag{6}$$

The oracle performs serial XOR, so the final target is $p_T = \bigoplus_{i=1}^{T} X_i$. At atomic step i, ψ_s incurs loss ℓ_{ok} when $X_i = s$ and ℓ_{err} otherwise; hence

$$R^\theta_{\mathcal{D}_i}(\psi_s) = \ell_{\mathrm{ok}} + \Delta_\ell P_\theta(X_i \neq s), \qquad \Delta_\ell = \ell_{\mathrm{err}} - \ell_{\mathrm{ok}} = \log 3.$$

Writing $s_\theta = (1+\theta)/2$, Eq. (6) gives

$$\sum_{i=1}^{T} \left[R^\theta_{\mathcal{D}_i}(\psi_{1-s_\theta}) - R^\theta_{\mathcal{D}_i}(\psi_{s_\theta})\right] = 2\Delta_\ell T\gamma. \tag{7}$$

The learned rollout uses the same s at every step. Since T is odd, its final hard output is s, so the final loss is ℓ_{ok} when $p_T = s$ and ℓ_{err} otherwise, and

$$R^\theta_{\mathcal{D}}(F_{\psi_{1-s_\theta}}) - R^\theta_{\mathcal{D}}(F_{\psi_{s_\theta}}) = 2\Delta_\ell T\gamma. \tag{8}$$

Equations (7) and (8) show that the atomic excess equals the final excess exactly.

For training, step i uses only the i-th coordinate of its own draws, written $X_{i1}, \ldots, X_{in}$, so the nT coordinates X_{ij} used are independent. The sub-task ERM therefore chooses

$$\widehat{s} = \mathbb{1}\left\{\sum_{i=1}^{T}\sum_{j=1}^{n} X_{ij} \geq \frac{nT}{2}\right\},$$

a majority test on nT independent Bernoulli variables with mean $1/2 + \theta\gamma$. The choice of γ guarantees $0 < \gamma \leq 1/(2T)$, and standard binomial anti-concentration gives a universal $c_1 > 0$ such that $P_\theta(\widehat{s} \neq s_\theta) \geq c_1$. Whenever the wrong hypothesis is selected, Eq. (7) gives atomic excess $2\Delta_\ell T\gamma$. Therefore

$$\mathbb{E}_S\left[\sum_{i=1}^{T} R^\theta_{\mathcal{D}_i}(\psi_{\widehat{s}}) - \min_{\psi\in\Psi}\left\{\sum_{i=1}^{T} R^\theta_{\mathcal{D}_i}(\psi)\right\}\right] \geq 2c_1\Delta_\ell T\gamma = \Omega\left(\min\left\{1, \sqrt{\frac{T}{n}}\right\}\right).$$

By Eq. (8) the atomic excess equals the final-task excess here, so the same bound holds for $\mathbb{E}_S[R^\theta_{\mathcal{D}}(F_{\psi_{\widehat{s}}}) - \min_{\psi\in\Psi} R^\theta_{\mathcal{D}}(F_\psi)]$; it is $\Omega(\sqrt{T/n})$ when $T \leq n$. □

What the lower bound certifies. The quantity bounded below is exactly the sample-dependent atomic excess of the upper bound—the sub-task ERM's excess over the best in-class value—and it coincides with the final-task excess. Since the per-step best-in-class loss is $\Theta(1)$ ($\ell_{\mathrm{ok}} + \frac{1}{2}\Delta_\ell \approx 0.837$), the excess is identifiable in this instance. Hence the $\Omega(\sqrt{T/n})$ rate is not an artifact of estimation: it concerns the estimation component only and says nothing about the approximation term.

C Full Experiments

In this section, we provide the extended experimental details and additional results that complement the main text.

C.1 Atom-Task Experiments

C.1.1 Experimental Setup

Atom-Task Benchmark. To validate compositional reasoning capability, we introduce Atom-Task. This task requires the model to sequentially execute a series of atomic operations on a symbolic sequence multiple times, necessitating multi-step reasoning capabilities.

Symbol System and Atomic Operations. The input consists of sequences formed by 6 Greek letters $\{\alpha, \beta, \gamma, \delta, \varepsilon, \omega\}$. We define five basic operations (mapped to IDs t1–t5):

- **Reverse (t1)**: Reverses the order of the sequence.
- **Substitute (t2)**: Swaps symbols in pairs ($\alpha \leftrightarrow \beta, \gamma \leftrightarrow \delta, \varepsilon \leftrightarrow \omega$).
- **Duplicate (t3)**: Duplicates each symbol consecutively.
- **Deduplicate (t4)**: Removes consecutive duplicates.
- **Rotate (t5)**: Moves the first element to the end.

Dataset Design and Distribution. A core challenge in compositional generalization is the distribution shift between training and testing. We design specific training datasets to investigate these effects.

Design Principles. Our training sets adhere to two core principles:

- **Dominance of atomic primitives**: In Mixed 1-, 3-, and 5-Step, single-step samples constitute 80% of the training trajectories, supporting the acquisition of basic operations.
- **Controlled exposure to composition**: Three- and five-step training chains provide structural patterns for compositional decomposition; in Mixed 3+5, the five-step chains use ≤ 2 distinct operations, as specified in Section 5.

Training Datasets. To evaluate OOD generalization, we design four training datasets:

1. **Single-Step**: 5k uniformly distributed one-step samples.
2. **Three-Step**: 1.66k three-step samples.
3. **Mixed 3+5**: 1,333 samples (833 three-step and 500 five-step; the five-step chains use at most two distinct operations, bridging length and combinatorial generalization).
4. **Mixed 1-, 3-, and 5-Step**: For each operation *op*, a dedicated 5k-sample dataset excludes *op* from all multi-step chains but keeps it in single-step data. Specifically, it contains 4,000 one-step, 600 three-step, and 400 five-step samples (5,000 training trajectories, or 7,800 atomic transitions in total; see Table 5).

Balanced Supervision: While the sample sizes vary across settings, the number of atomic transitions is held roughly constant: the single-step, three-step and mixed 3+5 datasets each contain about 5,000 transitions ($5{,}000$, $1{,}666 \times 3$, and $833 \times 3 + 500 \times 5$), while the operation-exclusion dataset contains $7{,}800$ and is compared across excluded operations under that same budget, which controls for confounding variables.

Test Sets. We evaluate on four test sets with increasing chain lengths: 3-step, 5-step, 7-step, and 9-step. Each set contains 500 unseen samples, requiring the model to perform compositional reasoning not explicitly memorized from training.

Table 5: Sample distribution in Ablation Datasets. Each dataset excludes a specific operation from multi-step chains. The 5,000 trajectories comprise 7,800 atomic transitions.

Data	Total	Single-step	3-step	5-step
`Ablation_no_*`	5,000	4,000	600	400

C.1.2 Implementation Details

For the Atom-Task, we use Qwen3-4B as the base model. Fine-tuning is performed using low-rank adaptation (LoRA; rank 16, alpha 32) for 3 epochs with a learning rate of 5×10^{-5}. The maximum generation length (`max_new_tokens`) is set to 2048 to accommodate longer reasoning chains. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Parameter Efficiency: We employ LoRA with the following configuration:

- Rank $r = 16$ for a balance between expressiveness and parameter efficiency.
- LoRA modules are applied to all linear layers within the attention and MLP blocks.

We employ a Chain-of-Thought training format, where the model is trained to explicitly predict the intermediate output after each atomic operation. For the "No-CoT" baseline, the model is trained to predict only the final output without intermediate states.

Reproducibility: For the main results (Table 1), both CoT and No-CoT entries are means across 3 random seeds (42, 2026, 12345), and the same 3 seeds are used for the operation-exclusion study, the format ablation, the noise tests, and the scaling points. For the format ablation (Table 3), the ablated CoT entries are means over 3 seeds and the original entries average the three-seed means of Table 2. Standard deviations are reported as subscripts and the seed setting is stated in each caption; differences of a few points between entries sharing a seed set should be read as indicative rather than as significance statements.

C.1.3 Details on Scaling Laws

To complement Figure 3, we provide the detailed accuracy numbers for training-sample scaling and inference-example scaling in Tables 6 and 7, respectively. All scaling experiments are conducted using the "Three-Step" training dataset and evaluated on held-out test sets, each reported point being a mean over 3 seeds (42, 2026, 12345).

Table 6: Accuracy (%) for Sample Size Scaling (n). Each entry is a mean over 3 seeds (42, 2026, 12345).

Prop.	3-step	5-step	7-step
10%	$91.93_{\pm 1.32}$	$78.00_{\pm 1.18}$	$58.73_{\pm 2.81}$
20%	$95.33_{\pm 0.41}$	$84.27_{\pm 1.27}$	$65.87_{\pm 2.33}$
50%	$98.13_{\pm 0.38}$	$91.40_{\pm 1.30}$	$80.40_{\pm 3.56}$
100%	$99.33_{\pm 0.25}$	$95.73_{\pm 0.68}$	$88.87_{\pm 1.54}$

Table 7: Accuracy (%) versus Number of Inference-Time Examples. Each entry is a mean over 3 seeds (42, 2026, 12345).

Shots	3-step	5-step	7-step
0-shot	$99.27_{\pm 0.09}$	$93.40_{\pm 2.83}$	$82.53_{\pm 1.76}$
1-shot	$99.33_{\pm 0.25}$	$95.73_{\pm 0.68}$	$88.87_{\pm 1.54}$
2-shot	$99.33_{\pm 0.09}$	$96.60_{\pm 0.16}$	$90.93_{\pm 1.32}$
3-shot	$99.07_{\pm 0.19}$	$95.73_{\pm 0.66}$	$89.93_{\pm 0.81}$

C.1.4 Extended Results and Analysis

Robustness to Noise in Reasoning Steps. To evaluate the robustness of CoT training to data quality, we conduct a noise interference experiment on the Qwen3-4B model. We test: **Multi-step (Pure 3-step)**, **Multi-step (Mixed 3+5)**, and the **w/o Substitute** ablation setting. (Note: These earlier noise robustness experiments used `max_new_tokens=1024`, but the conclusions remain consistent).

Noise Injection Strategy. We define the basic unit of noise injection as a single inference step. We select a proportion $p \in \{0.05, 0.1, 0.2\}$ of the total reasoning steps in the training set for interference. A standard reasoning step contains four discrete elements: the operation, the input state, the output state, and the remaining tasks. For each selected step, we randomly choose **two** of these four elements and perform either *Tampering* (replacing content with random symbols) or *Deletion* (emptying the content).

Table 8: Each entry is a mean over 3 seeds (42, 2026, 12345). Robustness to training noise (Qwen3-4B). Accuracy (%) under varying noise ratios. Each cell shows the result for the 3-step / 5-step / 7-step test sets.

Noise Ratio	Multi-step (Pure 3-step)	Multi-step (Mix$_{3+5}$)	w/o Substitute
$p = 0.05$	$99.73_{\pm 0.09}$ / $97.80_{\pm 0.86}$ / $90.47_{\pm 2.46}$	$99.33_{\pm 0.09}$ / $98.53_{\pm 0.34}$ / $94.87_{\pm 0.81}$	$96.20_{\pm 0.65}$ / $92.33_{\pm 1.06}$ / $85.47_{\pm 2.80}$
$p = 0.10$	$99.47_{\pm 0.09}$ / $97.20_{\pm 0.59}$ / $91.67_{\pm 0.25}$	$99.27_{\pm 0.34}$ / $98.87_{\pm 0.25}$ / $94.80_{\pm 0.49}$	$95.80_{\pm 0.33}$ / $92.33_{\pm 1.27}$ / $86.73_{\pm 1.04}$
$p = 0.20$	$99.40_{\pm 0.00}$ / $96.60_{\pm 0.75}$ / $90.73_{\pm 1.89}$	$99.13_{\pm 0.09}$ / $98.93_{\pm 0.62}$ / $94.80_{\pm 0.28}$	$96.93_{\pm 0.77}$ / $92.47_{\pm 0.50}$ / $87.13_{\pm 1.97}$

Table 8 presents accuracy for 3-step / 5-step / 7-step test sets respectively. The results indicate strong robustness to noise in reasoning steps. Across all training configurations, increasing the noise ratio does not lead to significant performance degradation; accuracy remains largely consistent with the clean baselines reported in previous tables. In some cases, minor variations are observed but without a clear negative trend, suggesting that the learned compositional reasoning patterns are resilient to localized inconsistencies in the training data.

Table 9: GSM8K baseline and comparative results. Each entry is a mean over 3 seeds (42, 2026, 12345).

Model	Experiment	Acc(%)
Llama2-7B	Untrained	2.50
	CoT Training	$39.14_{\pm 0.99}$
	No-CoT Training	$8.16_{\pm 2.50}$
Qwen2.5-3B	Untrained	64.44
	CoT Training	$79.23_{\pm 0.25}$
	No-CoT Training	$19.41_{\pm 1.17}$

Table 10: GSM8K accuracy under distribution shifts. Each entry is a mean over 3 seeds (42, 2026, 12345).

Model	Shift Type	Acc(%)
Llama2-7B	Scale Shift	$38.74_{\pm 0.70}$
	Complexity Shift	$36.57_{\pm 0.17}$
	Topic Shift	$37.51_{\pm 2.14}$
Qwen2.5-3B	Topic Shift	$78.60_{\pm 0.55}$
	Complexity Shift	$79.53_{\pm 0.45}$
	Answer Scale Shift	$79.10_{\pm 0.25}$

C.2 GSM8K Experiments

To validate our findings on real-world, complex reasoning scenarios, we conduct experiments on the GSM8K benchmark using Llama2-7B and Qwen2.5-3B fine-tuned on MetaMathQA, a math instruction-tuning dataset (395K samples in total). This protocol evaluates a controlled within-family distribution shift from MetaMathQA to GSM8K rather than source-disjoint transfer. Due to computational constraints, we use 30,000 training samples for each experiment. For standard experiments (baseline and noise), we randomly sample these 30,000 instances from the full dataset; for distribution shift experiments, we select 30,000 samples according to specific shift criteria.

C.2.1 Experimental Setup

Models and Training. We employ Llama2-7B and Qwen2.5-3B as base models. Models are fine-tuned using LoRA (rank=16, alpha=16) for 1 epoch with a learning rate of 2×10^{-5} and a maximum sequence length of 512. Training data is drawn from the MetaMathQA dataset. For baseline and noise experiments, we randomly sample 30,000 instances; for distribution shift experiments, we construct 30,000-sample subsets by filtering according to three shift indicators (Topic, Complexity, Answer Scale). All experiments are conducted on a single NVIDIA RTX 4090 GPU, and each experiment is run with three random seeds (42, 2026, 12345); the training data is identical across seeds, so reported standard deviations reflect model-side randomness.

Evaluation. All models are evaluated on the standard GSM8K test set. We report accuracy as the primary metric.

C.2.2 CoT vs. No-CoT & OOD Generalization

Baseline Comparison. As shown in Table 9, CoT training significantly boosts accuracy over untrained baselines, while No-CoT training yields only marginal improvements or even degradation.

Distribution Shift Experiments. In contrast to the random sampling used for baseline and noise studies, here we deliberately create distribution gaps by filtering 30,000 samples from MetaMathQA based on the following criteria:

- *Topic Shift (ood_domain)*: $\text{shift}_{\text{topic}} = \frac{|K_{\text{sample}} \setminus K_{\text{GSM8K}}|}{|K_{\text{sample}} \cup K_{\text{GSM8K}}|}$, where K_{sample} is the keyword set of the sample and K_{GSM8K} the top-20 keyword set of GSM8K; this asymmetric keyword-novelty score measures the fraction of sample keywords absent from GSM8K's set, normalized by the union size.
- *Complexity Shift (ood_complexity)*: $\text{shift}_{\text{complexity}} = \frac{|\text{steps}_{\text{sample}} - \mu_{\text{GSM8K}}|}{\sigma_{\text{GSM8K}}}$ (absolute Z-score of the sample's step count relative to GSM8K's mean and standard deviation).
- *Answer Scale Shift (ood_magnitude)*: $\text{shift}_{\text{magnitude}} = 1 - P_{\text{GSM8K}}(\text{magnitude}_{\text{sample}})$ (one minus the GSM8K probability of the sample's answer magnitude bin).

The results in Table 10 show that performance remains robust across all shifts, with only mild fluctuations. This stability suggests that the compositional patterns learned via CoT can generalize to practical domains like mathematical reasoning.

Robustness to Reasoning Noise. We further test resilience to imperfect supervision by training on 30,000 randomly sampled CoT instances from MetaMathQA, injected with noise via a controlled strategy: (1) *Sample-level probability*: Each sample has a probability p of being selected; (2) *Single-step injection*: For selected samples, noise is injected into a single randomly chosen reasoning step; (3) *Noise type*: 'delete' (removing the step content) or 'wrong_number' (modifying the numerical result), applied with equal probability (50%). This assesses the model's ability to handle corrupted intermediate reasoning.

Table 11: Each entry is a mean over 3 seeds (42, 2026, 12345). GSM8K accuracy under training with different proportions (p) of reasoning noise. All models trained on 30,000 samples.

Model	$p = 0$ **(Baseline)**	$p = 0.05$	$p = 0.10$	$p = 0.20$	$p = 0.50$
Llama2-7B	$39.14_{\pm 0.99}$	$39.20_{\pm 0.56}$	$39.73_{\pm 1.02}$	$37.63_{\pm 0.88}$	$37.66_{\pm 0.89}$
Qwen2.5-3B	$79.23_{\pm 0.25}$	$79.27_{\pm 0.25}$	$78.57_{\pm 0.37}$	$77.20_{\pm 0.60}$	$75.69_{\pm 0.56}$

Both models exhibit strong resistance to noise (Table 11). Even with up to 50% of samples corrupted, accuracy fluctuates within a narrow band and never collapses. This suggests that the model is relatively insensitive to localized noise in intermediate steps and can still absorb the correct logical patterns.

C.2.3 Summary of Findings

The GSM8K experiments reveal three main patterns: (1) CoT training substantially improves accuracy over untrained and No-CoT baselines, highlighting the benefit of intermediate step supervision; (2) models are robust to significant reasoning noise, maintaining stable performance even when training data contains deleted or incorrect steps, indicating that supervision remains effective even when individual steps are corrupted; (3) models generalize effectively across distribution shifts in topic, complexity, and answer scale, with only mild accuracy fluctuations, indicating that the reasoning patterns acquired via CoT transfer robustly to real-world mathematical domains.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: *The abstract and introduction clearly state the theoretical contributions (e.g., OOD generalization bounds via Rademacher complexity, error propagation coefficient independent of T, NTP-specific bound) and scope (studying CoT as compositional OOD generalization under a supervised NTP-style training paradigm). These claims are matched by the theoretical results of Sections 3–4 (Theorems 1, 2 and 5 and their corollaries, together with the tightness constructions of Appendix B) and by the experiments (Atom-Task, GSM8K).*

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: *A dedicated "Limitations & Future Work" paragraph in Section 7 explicitly states that the theory studies explicit CoT under supervised NTP-style training and excludes adaptive reasoning paradigms like reinforcement fine-tuning and test-time search.*

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: *Both assumptions are stated explicitly as Assumptions 1–2. Complete proofs for all theorems are provided in Appendix B; the main text states the results together with remarks on their implications and on their sharpness.*

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: *The paper details the experimental setup, including models (Qwen3-4B on Atom-Task; Llama2-7B and Qwen2.5-3B on GSM8K), training method (LoRA with rank=16), hyperparameters (learning rate, epochs), dataset design (Atom-Task operations, sample sizes), and evaluation protocol. Full details are in Appendix C.*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: *The abstract links the public code repository that reproduces the experiments.*

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: *Training and test details are specified in the main text (Section 5) and Appendix C (e.g., data splits, sample sizes, GPU type, LoRA settings, seeds).*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: *All result tables report means over three seeds (42, 2026, 12345), with standard deviations given as subscripts.*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: *Appendix C.2.1 states that "All experiments are conducted on a single NVIDIA RTX 4090 GPU."*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: *The research is theoretical and uses standard, publicly available models and synthetic/semi-synthetic benchmarks. There are no obvious violations of the Code of Ethics regarding human subjects, privacy, or fairness.*

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: *This is foundational theoretical and experimental research aimed at understanding the generalization mechanisms of CoT. It does not present a specific new application or deployed system with direct societal impacts at this stage. The paper focuses on the mathematical properties of an existing technique.*

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: *The paper does not release a new model or a high-risk dataset. It fine-tunes existing open-source models (Qwen, Llama) on a standard math dataset (MetaMathQA) and a novel symbolic benchmark (Atom-Task). These assets do not pose a high risk for misuse beyond what is already inherent in the base models.*

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: *The paper cites the original works for all used assets. The licenses are: MetaMathQA – MIT; Llama2-7B – LLAMA 2 COMMUNITY LICENSE; Qwen3-4B – Apache 2.0; GSM8K – MIT. All terms of use are respected.*

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: *The new "Atom-Task" benchmark is introduced and its operations (Reverse, Substitute, etc.) are clearly documented in Section 5 and Appendix C.1.1.*

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: *The paper does not involve crowdsourcing or research with human subjects.*

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: *The paper does not involve crowdsourcing or research with human subjects.*

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: *LLMs are the object of study rather than a component of the proposed method: our contribution is a theoretical decomposition of CoT error together with its empirical validation on Atom-Task and GSM8K, and no LLM module enters the method, the analysis, or the experimental pipeline as a non-standard component. LLM-based tools were used only for language editing of the manuscript, which under the policy above does not require declaration.*

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.