LEARNING WITHOUT OVERWRITING: A THEORY OF SELF-DISTILLATION AND SUPERVISED FINE-TUNING IN CONTINUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) of large language models (LLMs) has demonstrated the ability to improve reasoning capabilities while preserving previously acquired knowledge. Despite substantial empirical success, the dynamics of OPSD in continual reasoning remain incompletely understood. Modeling LLM reasoning as search over a directed acyclic graph, we provide a unified theoretical analysis of both the dynamics of post-training—OPSD and supervised fine-tuning (SFT) in continual learning—and the impact of pre-training on subsequent performance. Our findings establish three key insights with an optimization guarantee: (i) OPSD with hints from correct outputs enables continual learning without forgetting by sparse yet effective gradient descent updates induced by the hint structure. (ii) SFT on correct reasoning paths can lead to catastrophic forgetting due to dense updates along the training paths, which overwrite the information previously acquired. (iii) Diversity in pre-training is crucial for enabling a post-trained model to reach a correct output when a rollout starts from an intermediate state. Our results, supported by theoretical analysis, show that reliable continual reasoning depends on how post-training updates interact with the reasoning structure established during pre-training.

1 INTRODUCTION

Post-training has substantially improved the task-solving capabilities of large language models through supervised fine-tuning (Chung et al., 2022; Luo et al., 2025; Zhang et al., 2024) and reinforcement fine-tuning (Guo et al., 2025; Jin et al., 2025; Zeng et al., 2025). A central distinction between common implementations lies in the source of training trajectories: off-policy SFT learns from fixed demonstrations (Zhao et al., 2023; Zhang et al., 2023), whereas on-policy reinforcement learning (RL) uses reward feedback on trajectories sampled from the current model (Wang et al., 2023d; Zhao et al., 2026a). In the latter, the generation distribution determines which experiences are encountered during training, while itself evolving through subsequent updates (Kaufmann et al., 2025; Li et al., 2026). The distribution inherited from pre-training therefore influences which reasoning paths are sampled: even models with the same initial task accuracy can distribute probability differently across paths and risk missing some useful reasoning steps. Using a Markovian model of reasoning, Bu et al. (2025) showed that RL fine-tuning can favor high-probability paths while suppressing rare but useful ones. For SFT, preserving existing capabilities is a central challenge: sequential fine-tuning can cause catastrophic forgetting of earlier tasks and degrade general capabilities (Wang et al., 2023b; Li et al., 2024). Efforts to mitigate this problem include learning successive tasks in orthogonal low-rank subspaces (Wang et al., 2023a) and applying sharpness-aware minimization to favor flatter loss landscapes (Li et al., 2024). Furthermore, in code instruction tuning, Biderman et al. (2024) reported both reduced forgetting and better preservation of output diversity with LoRA than with full fine-tuning.

Self-distillation (SD) has also been explored as a means of learning new tasks while preserving previously acquired capabilities. In recent on-policy variants, the student generates a trajectory, and the same model—conditioned on an expert demonstration or informative feedback unavailable to the deployed student—acts as a teacher. By deriving supervision from the model's own behavior, self-distillation can reduce forgetting during adaptation (Yang et al., 2024; Shenfeld et al., 2026a). They

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

also provide token-level supervision on student rollouts, with reported improvements in reasoning accuracy over off-policy SFT and training efficiency over GRPO (Hübotter et al., 2026; Zhao et al., 2026b). These methods leverage the pre-trained model's in-context learning abilities (Shenfeld et al., 2026a; Hübotter et al., 2026), although SD can underperform RL in some settings (Yang et al., 2026; Kim et al., 2026). Their reliance on student trajectories also makes the diversity of the pre-trained policy relevant to learning.

Despite empirical successes, the dynamics of SD and their dependence on the pre-trained policy remain only partially understood. Wang et al. (2026a), the general analysis of self-improvement, established high-probability guarantees for true-risk improvement and finite-sample generalization under self-modification. Existing analyses for self-distillation characterize specific aspects of post-training. Using a Gaussian mixture model, Balasubramanian & Kasiviswanathan (2026) showed that SFT on new-task data alone drives the mixture weight associated with previously learned behavior to zero. For autoregressive models, Sang et al. (2026) bounded policy drift within each fixed-teacher window in terms of the distillation loss and the shift induced by a conciseness instruction. However, these analyses directly assume the teacher distribution rather than derive it from hint information. They consequently leave unresolved how hints—unique to self-distillation—shape learning dynamics, which is more difficult because hint-based analysis requires observations of the teacher behavior and its effect on the evolving student policy through multi-step reasoning.

Our objective is, therefore, to explain the difference in forgetting between on-policy SD and off-policy SFT, while considering their ability to learn new tasks. We model Chain-of-Thought (CoT) reasoning as a problem of reaching a task-specific goal in a DAG. This model links the sparsity of hint information to localized updates and retention, and provides a quantitative comparison with SFT. We evaluate this comparison numerically in Section 4. Although existing studies suggest that pre-training is important for the successful SD, there are few theoretical analyses on its role in on-policy SD. Hence, as a complementary theoretical discussion, we examine in Section 3.3 how limited diversity in the pre-training can leave relevant transitions unlearned and thereby impair reasoning from intermediate states.

Summary of contributions Our contributions are summarized as follows:

- We model the continual learning scenario in Section 2 with a graph-based approach, which is often used to explain LLM reasoning process such as CoT reasoning (Abbe et al. (2024); Yang et al. (2025); Kim et al. (2025)). In particular, we introduce a path-conflict setting to analyze the interference between tasks.
- We prove the learnability of a single task under both SD and SFT in Section 3.1 (Theorems 1 and 2, respectively). Notably, Lemma 4 in Appendix B.2 shows that SD with a hint in Equation 9 can realize a sparse update that preserves the performance of the pre-trained model. Furthermore, we show that SD can learn new tasks while preserving previously acquired capabilities, whereas SFT can forget them in Theorem 3 and Theorem 4. These results are aligned with the empirical findings of Shenfeld et al. (2026a).
- We conduct numerical experiments in Section 4 to support our theoretical bounds. The results confirm that tracing the hint-based sparse updates of SD and the overwriting properties of SFT is helpful to explain the observed learning dynamics.
- As a complementary theoretical analysis, we study the effect of pre-training diversity on reasoning from intermediate states in Section 3.3 (Proposition 1). We show that limited coverage in the pre-trained policy can leave relevant transitions unlearned even after SD succeeds from the original start.

These results explain how the sparsity of hint information translates into localized updates and retention in continual reasoning. The complementary pre-training analysis gives a theoretical implication for reasoning from intermediate states.

1.1 Related work

In the following, we briefly discuss related work on self-distillation and diversity in LLM reasoning. Additional related work are discussed in Appendix A.

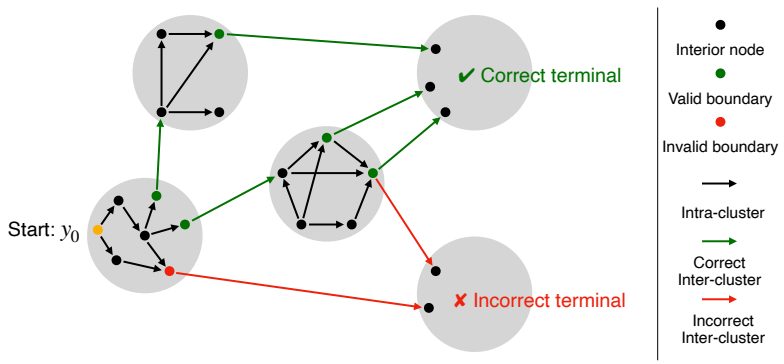

Figure 1: Example of the CoT graph we considered in Section 2. Each node represents a reasoning state, and edges correspond to reasoning transitions. Black edges indicate easy intra-cluster transitions, while red and green ones mean difficult inter-cluster transitions. The goal of training is to improve difficult transitions and to reach the correct terminal. Therefore, the model is required to learn the inter-cluster transitions while preserving the diversity within clusters.

Self-distillation for LLMs. Recent methods differ in the information supplied to the self-teacher. Shenfeld et al. (2026a) introduced SDFT, conditioning the teacher on pre-existing expert demonstrations to supervise student-generated trajectories. Their experiments demonstrated new-task acquisition with reduced forgetting relative to SFT. Hübotter et al. (2026) instead conditioned the SDPO teacher on feedback obtained after evaluating the student's own attempts, such as execution errors, improving reasoning accuracy and training efficiency. Zhao et al. (2026b) proposed OPSD, providing the teacher with reference answers or verified reasoning traces, and demonstrated improved mathematical reasoning with greater token efficiency than GRPO. Ye et al. (2026) introduced OPCD, including a self-distillation setting, using knowledge extracted from past solution traces or optimized system prompts as teacher context. Despite substantial empirical evidence, the learning dynamics of SD remain theoretically underexplored.

Diversity in LLM reasoning. Wang et al. (2023c) demonstrated that aggregating answers across diverse reasoning paths improves reasoning accuracy with numerical experiments. Yao et al. (2025) proposed diversity-aware policy optimization, which promotes token-level diversity selectively on correct responses. Their experiments showed improvements in both solution diversity and performance on mathematical reasoning benchmarks. Using a Markovian model, Bu et al. (2025) theoretically showed that post-training can suppress rare but useful reasoning paths and emphasize the role of exploration in preserving rare but useful reasoning paths already present in the pre-trained model.

2 Problem Setup

2.1 Continual learning on CoT graph

Following the reasoning-graph construction of Kim et al. (2025), we model the LLM chain-of-thought process as Markov transitions on a directed acyclic graph (**DAG**). Let the LLM CoT graph be $G = (V, E)$, where V is the node set and E is the directed edge set. A model receives input $x \in \mathbb{R}^d$ and generates a CoT sequence $y = (y_0, \ldots, y_t, \ldots, y_T)$, where $y_t \in V$, $(y_t, y_{t+1}) \in E$, and T denotes a rollout-dependent termination time. Note that y is node-simple since the graph is acyclic. We consider a correspondence between these states and the graph nodes; that is, each node in the graph is regarded as a reasoning state y_t like a logical sentence or a mathematical expression rather than a simple token. Thus, the prediction goal is to reach a proper node or cluster through logical reasoning transitions. The graph G consists of K clusters denoted by $C_1 \ldots C_K$, and it is free from multiple edges. A cluster is a set of similar reasoning states like equivalent reformulations obtained by straightforward manipulation. Hence, we consider the inter-cluster edges represent the essential and conceptual transitions, while intra-cluster edges are something trivial. For every node $u \in C_k$ and every cluster $C_{k'}$ with $k' \neq k$, there is at most one node $v \in C_{k'}$ such that $(u, v) \in E$.

We separate the cluster nodes in C_k into two types: one is the interior nodes denoted by $\mathring{C}_k$, and the other is the boundary nodes denoted by ∂C_k

$$\mathring{C}_k := \{u \in C_k : \forall v \in V, (u, v) \in E \Rightarrow v \in C_k\}, \quad \partial C_k := C_k \backslash \mathring{C}_k. \tag{1}$$

For preserving the inner-cluster transition diversity, we assume that all destinations of inter-cluster edges are interior nodes.

Continual learning. We study continual learning of L tasks $\mathcal{D}_1, \ldots, \mathcal{D}_L$. The task $\mathcal{D}_\ell$ consists of a fixed input $x_\ell \in \{0,1\}^d$, a fixed CoT startpoint $y_{\ell,0} \in \mathring{C}_{k_\ell}$, and a task-specific correct terminal cluster $\tilde{C}_\ell \in \{C_1, \ldots, C_K\}$. We assume that $\tilde{C}_\ell$ has no outgoing inter-cluster transitions. Furthermore, we assume that the tasks can be separated. More precisely, the matrix $(x_1, \ldots, x_L)$ is column full rank and the clusters $\tilde{C}_1, \ldots, \tilde{C}_L$ are disjoint. A rollout for task $\mathcal{D}_\ell$ is defined to be correct if and only if $y_T \in \tilde{C}_\ell$. [1] Then, we consider the following assumption.

Assumption 1 (transition uniqueness and competition between tasks)**.** *Each boundary node has at most one correct cluster transition for each task. Whenever such a transition exists for multiple tasks at the same boundary node, it is identical across any tasks that are not orthogonal in the input. Hence, competition between tasks can arise only through trivial interior transitions.*

This assumption is strong, but seems reasonable since similar tasks share underlying reasoning properties, which we model as common inter-cluster transitions. $\mathcal{P}_\ell$ is the set of correct paths for task $\mathcal{D}_\ell$, and E_ℓ is the union of the inter-cluster transitions appearing in those correct paths. In learning, the model is trained sequentially on the tasks $\mathcal{D}_1, \ldots, \mathcal{D}_L$.

2.2 Pre-training and fine-tuning

Pre-training requirements For each task $\mathcal{D}_\ell$ and cluster C_k, define the set of boundary nodes that admit a correct inter-cluster transition by

$$\partial^\star_\ell C_k = \{u \in \partial C_k \mid \exists v \in V \backslash C_k, (u, v) \in E_\ell\}. \tag{2}$$

Then, we further define $M = \max_{k,\ell} |\partial^\star_\ell C_k|$. We call the initial node or an entry through a valid inter-cluster transition a relevant start s in C_k. For a rollout initialized at $Y_0 = s$ under task input x_ℓ, define

$$\tau^s_{\ell,k} := \inf\{t \geq 0 : Y_t \in \partial C_k\}, \quad U^s_{\ell,k} := Y_{\tau^s_{\ell,k}}.$$

Thus $U^s_{\ell,k}$ is the first boundary node reached after the rollout starts from the relevant start s. For the model, we consider a softmax predictor:

$$\pi_{\mathbf{W}}(\cdot \mid u, x) = \mathrm{softmax}(x^\top \mathbf{W} e_{u,\cdot}), \quad \mathbf{W} \in \mathbb{R}^{d \times |E|}, \tag{3}$$

where $e_{u,\cdot}$ is the embedded concatenation matrix of one-hot vectors $e_{u,v}$ for all v s.t. $(u, v) \in E$. The following assumption specifies the required properties of the pre-trained parameter $\mathbf{W}^{(0)}$.

Assumption 2 (Pre-trained diversity)**.** *For every task $\mathcal{D}_\ell$ and non-terminal cluster C_k satisfying $\partial^\star_\ell C_k \neq \emptyset$, the first boundary-hitting time after every relevant start s in C_k is almost surely finite. Uniformly over all such s, the pretrained model satisfies*

$$\mathbb{P}_{W^{(0)}}\left(U^s_{\ell,k} \notin \partial^\star_\ell C_k \mid Y_0 = s, x_\ell\right) = O(\epsilon_1), \tag{4}$$

$$\mathbb{P}_{W^{(0)}}\left(U^s_{\ell,k} = u \mid Y_0 = s, x_\ell\right) = \Omega(1/M), \tag{5}$$

for every $u \in \partial^\star_\ell C_k$. Moreover, $\pi_{W^{(0)}}(v \mid u, x_\ell) = \Theta(\epsilon_2)$ for every inter-cluster edge (u, v) and every task input x_ℓ. ϵ_1 and ϵ_2 are supposed to be sufficiently small; at least, they satisfy $KM\epsilon_1 = o(1)$ and $M\epsilon_2 = o(1)$.

The diversity conditions state that (i) the pre-trained model reaches valid boundaries with high probability, and (ii) every inter-cluster transition initially has small but nonzero probability. In other words, these assumptions suggest that the trivial reasoning steps are sufficiently acquired, but essential transitions are difficult and the model is required further training to learn. Appendix B.1 gives a finite-parameter satisfying every condition of Assumption 2 on a family of layered DAGs.

[1] We focus on reaching the terminal cluster. For the subsequent search for a particular target node within it, Kim et al. (2025, Lemma B.9) provide hitting-time bounds under additional assumptions on within-cluster mixing and stationary probabilities.

Fine-tuning settings We use SD and SFT as on- and off-policy fine-tuning methods respectively. The detailed algorithms are described in Algorithm 1. For each task $\mathcal{D}_\ell$, a teacher-forcing path $y^* = (y_0^*, \dots, y_T^*)$ and a hint c are constructed by sampling

$$\gamma \sim \mathrm{Unif}(\mathcal{P}_\ell). \tag{6}$$

Across training iterations and across elements within each mini-batch, the teacher paths are sampled independently according to Equation 6, conditional on the preceding training history. Note that for SD, conditional on the same history, each teacher path used to construct the hint and its corresponding unhinted student rollout are sampled independently. We define the teacher path to be $y^* = \gamma$, and the hint c represents all cluster transitions in γ; that is, the hint is $c = \{c_i\}_{i=1}^{I}$, where $c_i = \{(u_i, v_i) \text{ is traversed.}\}$. This hint construction reflects the human instruction that ignores trivial transitions and emphasizes the essential reasoning steps. At each $\mathcal{D}_\ell$ training iteration, the model receives a minibatch of $\{x_\ell, c^{(b)}\}_{b=1}^{B}$ or $\{x_\ell, y^{*(b)}\}_{b=1}^{B}$ for SD or SFT respectively, and updates the parameter $\mathbf{W}$ by minimizing the loss function. Note that such an external hint is consistent with Shenfeld et al. (2026a) and Liu et al. (2026). Since SD uses only hints about specific transitions rather than the more informative full paths used in SFT, pre-training is crucial for its success. We use the same uniforum distribution for SD and SFT to compare their performance under a common source of supervision without favoring any correct path. Batch size B is fixed for all tasks and iterations. The mini-batch loss functions for SD and SFT are defined as follows respectively:

$$L_{\mathrm{SD}}(\mathbf{W}; \mathcal{D}_\ell) = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=0}^{T^{(b)}-1} D_{\mathrm{KL}} \left(\pi_{\mathbf{W}}(\cdot \mid y_t^{(b)}, x_\ell) \parallel \mathrm{SG}(\pi_{\mathbf{W}}(\cdot \mid y_t^{(b)}, x_\ell, c^{(b)})) \right), \tag{7}$$

$$L_{\mathrm{SFT}}(\mathbf{W}; \mathcal{D}_\ell) = -\frac{1}{B} \sum_{b=1}^{B} \sum_{t=0}^{T^{(b)}-1} \log \pi_{\mathbf{W}}(y_{t+1}^{*(b)} \mid y_t^{*(b)}, x_\ell), \tag{8}$$

where SG denotes stop-gradient. Especially, we do not batch the KL-divergence to use the same SD loss as Shenfeld et al. (2026a) and Liu et al. (2026). Both SD and SFT update the parameters using SGD as $\mathbf{W}^{(n+1)} \to \mathbf{W}^{(n)} - \eta_j \nabla_{\mathbf{W}} L_{\mathrm{method}}(\mathbf{W}^{(n)}; \mathcal{D}_\ell)$, where $\mathrm{method} \in \{\mathrm{SD}, \mathrm{SFT}\}$ denotes the corresponding training method.

3 Main results

If the hint c strictly indicates that the model should follow all cluster shift (u_i, v_i) in γ, then the teacher support would be strictly subset of the student one, and the KL divergence would diverge. Avoiding this divergence, fix a smoothing parameter $\epsilon_{\mathrm{rel}} \in (0, 1)$ and use a relaxed hint c:

$$\chi_{\mathbf{W}}(c_i \mid y_{0:T}, x) \propto \begin{cases} 1 & (\tau_i < T, y_{\tau_i+1} = v_i), \\ \epsilon_{\mathrm{rel}} & (\tau_i < T, y_{\tau_i+1} \neq v_i), \\ Z_{\mathbf{W},i}(x) & (\tau_i \geq T), \end{cases} \tag{9}$$

where $\tau_i = \inf\{s \in \{0, \dots, T\} \mid y_s = u_i\}$, with the convention $\inf \emptyset = \infty$, and $Z_{\mathbf{W},i}(x) = \pi_{\mathbf{W}}(v_i \mid u_i, x) + \epsilon_{\mathrm{rel}}(1 - \pi_{\mathbf{W}}(v_i \mid u_i, x))$. $\chi_{\mathbf{W}}$ is a likelihood factor for the relaxed hint c_i, and $Z_{\mathbf{W},i}(x)$ is equivalent to the expected local feedback before the transition from u_i is observed. When u_i is not visited, using this expected value as feedback allows us to leave the corresponding transition parameters unchanged, since the current student provides no information about transition out of u_i. This edge-level feedback enables the sparse activation, which identifies where mistakes happen and adjusts to the student's response distribution (Hübotter et al. (2026)). From (9), the teacher distribution is determined by Bayes' rule:

$$\pi_{\mathbf{W}}(y_{0:T} \mid c, x) \propto \pi_{\mathbf{W}}(y_{0:T} \mid x) \cdot \chi_{\mathbf{W}}(c \mid y_{0:T}, x). \tag{10}$$

The step-level teacher distribution $\pi_{\mathbf{W}}(y_{t+1} \mid y_t, c, x)$ is derived in Appendix B.2.

Sparse structure learning. We believe that hint-induced sparse learning mitigates catastrophic forgetting. Indeed, as shown in Lemma 4 in Appendix B.2, the hint in Equation 9 yields a gradient that updates only the parameters associated with outgoing edges from the hint-source nodes. Consequently, SD preserves the interior transition diversity, which is crucial for continual learning, whereas SFT updates the parameters at all source nodes along the teacher path.

3.1 SINGLE-TASK LEARNABILITY

SD as on-policy training Before analyzing the continual learning performance, we first consider their ability to learn a single task. Theorems 1 and 2 suggest that SD and SFT are both effective for solving one task. It is important to note that there is little difference between the two methods in learning a single task.

Theorem 1 (Single-task learning under SD). *Fix a task $\mathcal{D}_\ell$ and let the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in [0, 1/2)$. Suppose Assumptions 1 and 2 hold and ϵ_1 is sufficiently small. Define*

$$\xi_{\ell,n}^{(K)}(\delta) := \left[\epsilon_2 + \epsilon_2^2 \left(\log \frac{1}{\epsilon_{\mathrm{rel}}}\right) \left[\frac{e^{-\Theta(K \log \frac{1}{\epsilon_2})} n^{1-\nu}}{M^K |\mathcal{P}_\ell|} - \sqrt{\frac{\log(KM/\delta)}{B}} n^{1/2-\nu}\right]^+\right]^{-1}. \quad (11)$$

Then, with probability at least $1-\delta$,

$$\pi_{\mathbf{W}^{(n)}}\left(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell\right) \geq 1 - O(KM\epsilon_1) - O\Big(KM\xi_{\ell,n}^{(K)}(\delta)\Big). \quad (12)$$

Theorem 1 says that SD can learn a single task with a success probability of $1 - O(KM\epsilon_1)$. The error term arises from visits to invalid boundaries, and in this setting, we reduce this error by assuming ϵ_1 is sufficiently small. The error $O(KM\epsilon_1)$ reflects a trade-off between single-task accuracy and the ability to retain previously acquired knowledge during continual learning. Preserving the student's pre-trained diversity also leaves some of its initial errors intact. On the other hand, this term might be improved by providing hints that discourage the model from proceeding in an incorrect direction. Indeed, Zhu et al. (2025) show that penalizing incorrect responses alone can improve reasoning performance, and their gradient analysis indicates that these updates suppress incorrect generations and redistribute probability mass toward plausible alternatives favored by the model's prior. During inference, Yao et al. (2023) use negative evaluations of intermediate reasoning states to prune unpromising branches in Tree of Thoughts, with ablation results showing that removing this pruning degrades search performance. Theorem 1 further implies that the strength of the hint controls the convergence rate of the student success probability. In fact, decreasing ϵ_{rel} reduces the bound in Equation 11 without increasing the number of iterations n. Thus, the theorem effectively captures how hint strength affects self-distillation dynamics.

SFT as off-policy learning Theorem 2 is a counterpart of Theorem 1 for SFT. SFT also learns a single task with high probability, and updates the parameters at all source nodes along the teacher. It is a main difference from SD, which only focuses on the source nodes related to the hint.

Theorem 2 (Single-task learning under SFT). *Fix $\delta \in (0, 1)$. Suppose the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in (0, 1/2)$ and $\mathbf{W}_{\mathrm{SFT}}^{(0)} = \mathbf{W}^{(0)}$. Then, with probability at least $1-\delta$,*

$$\min_{u \in \mathcal{U}_\ell} \pi_{\mathbf{W}_{\mathrm{SFT}}^{(n)}}(\mathcal{A}_\ell(u) \mid u, x_\ell) \geq 1 - \tilde{O}\left(|\mathcal{P}_\ell| \left[\frac{1}{n^{1-\nu}} + \frac{\log(e/\delta)}{Bn^\nu}\right]\right). \quad (13)$$

Here $\mathcal{U}_\ell$ is the set of nodes appearing in the correct paths except for the terminal, and $\mathcal{A}_\ell(u)$ is the set of their correct successors from u. $\tilde{O}(\cdot)$ hides logarithmic factors in n and $|\mathcal{P}_\ell|$.

Sketch of Proof. Note that the optimal parameter $\mathbf{W}_{\mathrm{opt}}$ is not finite because a softmax predictor with a finite parameter cannot assign probability one to any outcome. Thus, we analyze the convergence in two steps: (1) Set parameter $\mathbf{W}_R \to \mathbf{W}_{\mathrm{opt}}(R \to \infty)$, and use it as a reference. (2) Construct the potential function to measure the difference between the current parameter, the reference parameter and the optimal one. More concrete analysis is provided in Appendix B.5 □

Theorem 2 shows that SFT learns to select correct successors from states on correct paths. Whereas SD carries over some of the pre-trained model's routing errors, Theorem 2 show that SFT can learn to generate correct paths with probability approaching one. This difference reflects the scope of the updates. SFT supervises all transitions along the teacher paths, including intra-cluster transitions, and can therefore correct the pre-trained routing errors reflected in the residual term of the SD lower bound. In the next section, we use this overwriting property to explain why SFT can forget previously learned tasks.

3.2 CONTINUAL LEARNING AND CATASTROPHIC FORGETTING

Sustainability of SD In this section, we study the continual learning performance of SD and SFT. A key feature of SD in our framework is that hints localize parameter updates to a small subset of graph nodes while SFT fails and overwrites the parameters of the student model. Intuitively, such localized updates may help preserve existing knowledge and mitigate catastrophic forgetting. Relatedly, Wang et al. (2026b) suggest that self-distillation can mitigate knowledge forgetting by selectively encouraging exploration at reasoning tokens while maintaining stable targets for factual tokens. The following theorem supports our intuition.

Theorem 3 (No forgetting under SD)**.** *Let $\mathbf{W}_{\mathrm{SD}}^{[s]}$ denote the parameter after sequentially training $\mathcal{D}_1, \ldots, \mathcal{D}_s$ by SD. Fix $\ell \in \{1, \ldots, L\}$, and let n_ℓ be the number of SD updates used for $\mathcal{D}_\ell$. Suppose Assumptions 1 and 2 hold for all tasks, and the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in [0, 1/2)$. Then, after any numbers of subsequent SD updates on $\mathcal{D}_{\ell+1}, \ldots, \mathcal{D}_L$, with probability at least $1 - \delta$,*

$$\pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}\left(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell\right) \geq 1 - O(KM\epsilon_1) - O\Big(KM\xi_{\ell,n_\ell}^{(K)}(\delta)\Big). \tag{14}$$

Moreover, every correct transition learned for task $\mathcal{D}_\ell$ satisfies

$$\pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}(v \mid u, x_\ell) \geq \pi_{\mathbf{W}_{\mathrm{SD}}^{[\ell]}}(v \mid u, x_\ell), \qquad (u, v) \in E_\ell. \tag{15}$$

Theorem 3 shows that, after learning task $\mathcal{D}_\ell$, the student can continue learning subsequent tasks through self-distillation while retaining its ability to generate correct rollouts from $y_{\ell,0}$ (Equations 14 and 15). Equation 15 also shows that the probability of correct inter-cluster transitions may even increase by subsequent tasks when the tasks share the same correct inter-cluster transitions because the hints associated with these transitions are sampled as well during training on subsequent tasks.

Forgetting in SFT SFT, on the other hand, is a learning system that forces the student to follow the teacher path, which is a relatively dense graph structure. Therefore, if any path conflict exists between tasks, the student rollout is likely to follow new task paths and forget the old ones. *We consider this overwriting is precisely what causes catastrophic forgetting.* For the SFT comparison, fix two tasks $\mathcal{D}_{\mathrm{old}}$ and $\mathcal{D}_{\mathrm{new}}$ learned in this order. Let $\mathbf{W}_{\mathrm{new}}^{(0)}$ be the parameter immediately after learning $\mathcal{D}_{\mathrm{old}}$, and let $\mathbf{W}_{\mathrm{new}}^{(n)}$ be the parameter after n subsequent SFT updates on $\mathcal{D}_{\mathrm{new}}$. Let $\theta \in [0, 1]$ denote the fraction of old correct paths that conflict with the new task. Formal definitions are deferred to Appendix B.7.

Theorem 4 (Catastrophic forgetting under SFT)**.** *Fix $\delta \in (0, 1)$ and set $\lambda := \frac{\langle x_{\mathrm{old}}, x_{\mathrm{new}} \rangle}{\|x_{\mathrm{new}}\|_2^2} > 0$. Suppose $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in (0, 1/2)$. Then, for arbitrary pre-new-task old-path masses, with probability at least $1 - \delta$,*

$$\Pi(\mathbf{W}_{\mathrm{new}}^{(n)}) \leq \underbrace{\mathfrak{C}_{\mathrm{old}}(1-\theta)}_{\substack{\text{upper bound on} \\ \text{surviving knowledge}}} + \widetilde{O}\left(\theta |\mathcal{P}_{\mathrm{old}}| |\mathcal{P}_{\mathrm{new}}|^\lambda \left[\frac{1 + B^{-1}\log(1/\delta)}{n^\nu}\right]^\lambda\right), \tag{16}$$

$$\Pi(\mathbf{W}_{\mathrm{new}}^{(0)}) - \mathfrak{C}_{\mathrm{old}}(\theta) \leq \Pi(\mathbf{W}_{\mathrm{new}}^{(n)}), \tag{17}$$

where $\Pi(\mathbf{W}) = \pi_{\mathbf{W}}(Y_T \in \tilde{C}_{\mathrm{old}} \mid Y_0 = y_{\mathrm{old},0}, x_{\mathrm{old}})$ and

$$\mathfrak{C}_{\mathrm{old}}(\phi) := \max_{\substack{\mathcal{P} \subseteq \mathcal{P}_{\mathrm{old}} \\ |\mathcal{P}| \leq \phi |\mathcal{P}_{\mathrm{old}}|}} \sum_{\gamma \in \mathcal{P}} \pi_{\mathbf{W}_{\mathrm{new}}^{(0)}}(Y_{0:T} = \gamma \mid Y_0 = y_{\mathrm{old},0}, x_{\mathrm{old}}). \tag{18}$$

$\tilde{O}(\cdot)$ hides logarithmic factors in n and $|\mathcal{P}_{\mathrm{new}}|$.

During sequential learning, Theorem 4 shows that SFT drives the probability mass of old correct paths that conflict with the new task toward zero as training proceeds (Equation 16). Despite some forgetting, SFT retains previously acquired knowledge that does not conflict with the new knowledge (Equation 17). Although our current setting excludes conflicts between tasks at inter-cluster transitions, this result may help explain why SD could remain less prone to forgetting than SFT in broader settings that allow such conflicts. Whereas SFT updates transitions along teacher paths, SD restricts its updates only to source nodes selected by hints, which may leave more old correct paths unaffected by subsequent training. If these paths account for most of the old task's success probability at the start of new task training, their preserved solution paths would alleviate forgetting.

Remark 1 (Explicit representation of $\mathfrak{C}_{\mathrm{old}}(\phi)$)**.** *Let $\kappa^{\mathrm{old}}_{\max}$ and $\kappa^{\mathrm{old}}_{\min}$ be the largest and smallest old-path masses normalized by their average mass, as defined in Appendix B.7. Then*

$$\mathfrak{C}_{\mathrm{old}}(\phi) \le \Pi(\mathbf{W}^{(0)}_{\mathrm{new}}) \min\left\{\kappa^{\mathrm{old}}_{\max}\phi, 1-\kappa^{\mathrm{old}}_{\min}+\kappa^{\mathrm{old}}_{\min}\phi\right\}. \tag{19}$$

Remark 2 (Error accumulation across tasks)**.** *For multiple later tasks, the union of their conflicted old-path sets can increase, and the fraction of non-conflicted old correct paths can decrease. Consequently, θ_{total}, the multi-task counterpart of θ, can increase as the number of tasks grows, thereby exacerbating the forgetting effect and further decreasing the corresponding concentration term in the success probability of the old tasks. This suggests an asymptotic upper bound of $\mathfrak{C}_{\mathrm{old}}(1-\theta_{\mathrm{total}})$ on the old-task success probability.*

3.3 A THEORETICAL IMPLICATION FOR PRE-TRAINING DIVERSITY

The preceding results concern new-task learning and retention from each task's original start. As a complementary theoretical discussion, we now ask when the learned policy can also reason successfully from intermediate states. The role of the pre-trained model in SD (Shenfeld et al., 2026a; Hübotter et al., 2026) and the benefits of diverse reasoning paths (Wang et al., 2023c) motivate this question. We examine diversity through the coverage of reasoning states in the pre-trained policy: because SD updates depend on visited states, limited coverage can leave some transitions unlearned that are needed for intermediate restarts. We fix a task $\mathcal{D}_\ell$ and run the same SD procedure from its original start; restarts are used only for evaluation. Under the coverage condition of Assumption 2, every correct boundary receives training opportunities and they suffice to support all relevant starts simultaneously (Proposition 3 in Appendix B.8). However, if the pre-trained model is not diverse enough, some relevant starts may not pass through proper cluster transitions and thus the model cannot generate a correct path from them.

Assumption 3 (Existence of a rare entrance)**.** *For the fixed task, retain Assumption 2 except its coverage lower bound at one correct boundary node u^*. The pretrained probability of ever visiting u^* from the original start is $O(1/M')$, with $M' = \Omega(M)$ and a uniform implied constant. Its correct transition is the only incoming edge to a nonterminal cluster C_k that does not contain the original start $y_{\ell,0}$.*

Proposition 1 (Restart failure beyond a rare entrance)**.** *Suppose Assumptions 1 and 3 hold, $\eta_j = \Theta(j^{-\nu})$ for fixed $\nu \in [0, 1/2)$, and $0 < \epsilon_1, \epsilon_2, \epsilon_{\mathrm{rel}}, \delta < 1$. If*

$$n = \Omega\left(\max\left\{\frac{\log(2KM/\delta)}{B\underline{\rho}^2_{\ell,K,M}}\,,\ \left[\frac{\epsilon_1^{-1}-1}{\epsilon_2^2\log(1/\epsilon_{\mathrm{rel}})\underline{\rho}_{\ell,K,M}}\right]^{1/(1-\nu)}\right\}\right), \tag{20}$$

$$M' = \Omega\left(\frac{1}{\epsilon_2^{|E_\ell|}}\max\left\{\frac{1}{KM\epsilon_1}, \frac{Bn}{\delta}\right\}\right), \tag{21}$$

with sufficiently large uniform constants, then, with probability at least $1-\delta$, no training rollout visits u^*, and jointly

$$\pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) = 1 - O(KM\epsilon_1) \quad \text{(Misleading high accuracy)}, \tag{22}$$

$$\sup_{s\in\mathcal{S}_{\ell,k}} \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell) = O(M\epsilon_2) \quad \text{(Restart failure)}. \tag{23}$$

Here $\mathcal{S}_{\ell,k}$ contains the relevant starts in C_k and $\underline{\rho}_{\ell,K,M}$ is specified in Appendix B.3.

Proposition 1 states that if C_k is a difficult cluster for the student to reach and the amount of training is insufficient, the student may still have a low success probability when restarting from relevant intermediate states in C_k even though the model seems to have highly learned the task from its original start $y_{\ell,0}$ (Equation 22). Rare entrances satisfying Assumption 3 make some clusters severely difficult to reach from the original start, and thus leave these clusters unlearned, preventing the success probability from increasing. To overcome this issue, self-distillation requires an amount of training that grows at least linearly with M'. The gap in success probabilities between the original start and the downstream cluster suggests that limited diversity can pose risks at inference time that are difficult to detect during training. Improved performance from the original start may not translate into better performance when restarting from downstream clusters. Pretrained coverage therefore supports intermediate restartability beyond original-start accuracy.

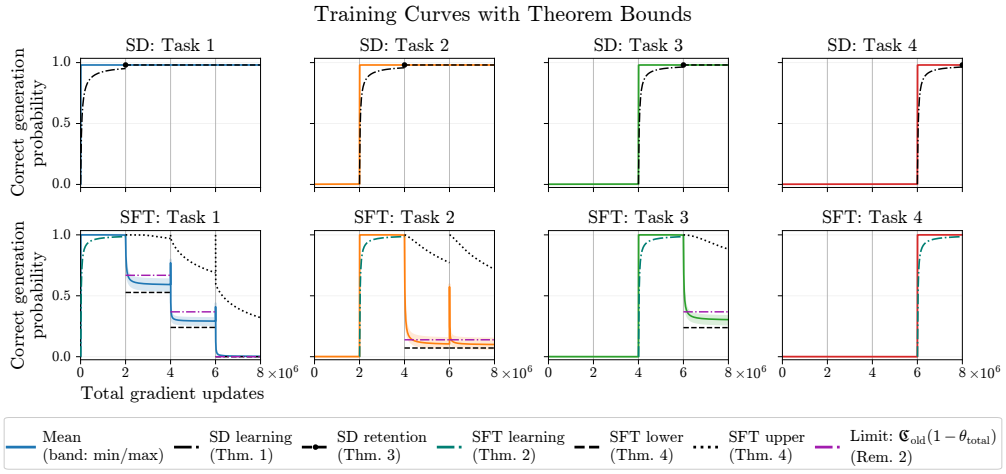

Figure 2: Training dynamics under SD (top row) and SFT (bottom row) for Tasks 1–4. Solid curves show the mean correct-generation probability over ten seeds, and shaded regions indicate the minimum–maximum range. Dash-dotted and dashed black curves represent the SD population learning references and conditional retention lower bounds. Dashed and dotted black curves represent the SFT conditional retention lower bounds and population upper references, respectively. Population references use $B \to \infty$ for stabilizing the curves. Purple dash-dotted lines show the concentration-profile term $\mathfrak{C}_{\text{old}}(1-\theta_{\text{total}})$. Vertical gray lines indicate task transitions.

Remark 3 (Separation of training times)**.** *Under Assumptions 1 and 2, the model trained by SD can resume from any relevant start at the same training length Equation 20 (Proposition 3 in Appendix B.8.) Holding other quantities fixed, the nonvisit horizon grows linearly with M'. If the nonvisit horizon grows asymptotically faster than the sufficient learning scale, there is a range of training lengths for which the two settings exhibit different restart behavior. Under the full-coverage condition, the trained model can reason successfully from every relevant start, whereas under the rare-entrance condition, the cluster beyond the rare entrance remains unlearned.*

4 Numerical Experiments

In this section, we evaluate whether the theoretical bounds are effective. The number of tasks is set to $L = 4$. Additional experimental details are written in Appendix C. Figure 2 shows training curves for every task between SD and SFT in continual learning. SFT exhibits a significant drop in performance on task 1 to 3 while SD preserves each task's performance during subsequent-task training. These results confirm that tracing the sparse updates in SD and the conflict between tasks in SFT can explain the observed retention and forgetting.

5 Conclusion

We analyze continual learning under SD and SFT using a CoT reasoning graph. Our theory shows how the sparsity of hint information localizes SD updates to preserve previously acquired capabilities, whereas dense SFT updates along the teacher paths can cause forgetting. As a complementary theoretical implication, we show that limited coverage in the pre-trained policy can prevent successful reasoning from intermediate states even after SD succeeds from the original start.

Future work. The major improvement is to extend the analysis from the current outcome reward format such as Equation 9 to that of the process reward. As shown in Hübotter et al. (2026), the self-teacher in on-policy self-distillation can act as an implicit process reward model through retrospection when provided with rich feedback. This setting might also be helpful to analyze the importance of in-context learning ability in the pre-training strategy.

AI USE STATEMENT

We used AI, mainly ChatGPT, to assist in the following ways.

- To aid and polish writing: We used AI to improve the clarity and grammar of the text, and to help us express our ideas more clearly.
- For retrieval of relevant references and related work: Mainly, we used AI to find and summarize relevant works and we incorporated them into this paper.
- For research ideation: We used AI to brainstorm ideas and to discuss the research direction. Assumption 1, Equations 4 and 5 in Assumption 2 and Section 3.3 is constructed during the discussion with AI.
- For the purpose of drafting sections of all contents in the paper: We made AI-generated drafts at all sections, which were then revised and modified repeatedly.
- For numerical experiments: We made AI create some code and the graph we used to validate our theoretical results. These codes and graphs were revised and modified by us to fit our purpose.
- For proving mathematical claims: AI helped us to check the correctness of the proofs and formalized our ideas to mathematical statements and provided proofs. In particular, the following is a list of the major works of AI in the proofs.
  - Lemma 9: AI helped us to evaluate the times of visitation to the correct boundaries.
  - Lemma 13: AI helped us to define the potential function in Equation 65 and then incorporated it into our original proof to produce the current version.
  - Theorem 4: AI helped us to prove the bounds in Theorem 4.

On the other hand, we did not use AI at all in the following points.

- **To create intuitive explanations of the results**: Theoretical frameworks like problem setting are not created by AI, but is constructed only based on our original ideas and previous related work.
- **To create the base ideas for proofs**: All mathematical claims such as theorems and propositions are based on our original proofs. Especially, in Theorems 1 and 3 and Proposition 1, no AI solutions are used to construct them.

REPRODUCIBILITY

Numerical experiments conducted in this paper are reproducible by using the code provided with this submission. The code includes the implementation of the synthetic CoT graph, the training procedures for SD and SFT. Experimental setting details are listed in Appendix C.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad, 2024. URL https://arxiv.org/abs/2406.06467.

Gregor Bachmann and Vaishnavh Nagarajan. The Pitfalls of Next-Token Prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2296–2318, 2024. URL https://proceedings.mlr.press/v235/bachmann24a.html.

Krishnakumar Balasubramanian and Shiva Prasad Kasiviswanathan. A quantitative characterization of forgetting in post-training, 2026. URL https://arxiv.org/abs/2603.12163.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA Learns Less and Forgets Less. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=aloEru2qCG.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Bo Xue, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. Distributional Biases in Post-Training: A Markovian Analysis of Reasoning Trajectories. *arXiv preprint arXiv:2511.07368*, 2025. URL https://arxiv.org/abs/2511.07368. Preprint.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025. URL https://arxiv.org/abs/2510.18874.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen

Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.20802.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2312.14925.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs?, 2026. URL https://arxiv.org/abs/2603.24472.

Juno Kim, Denny Wu, Jason D. Lee, and Taiji Suzuki. Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 30791–30825, 2025. URL https://proceedings.mlr.press/v267/kim25aj.html.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *arXiv preprint arXiv:2507.05386*, 2025. URL https://arxiv.org/abs/2507.05386.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting Catastrophic Forgetting in Large Language Model Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.249. URL https://aclanthology.org/2024.findings-emnlp.249/.

Ziniu Li, Pengyuan Wang, Tian Xu, Tian Ding, Ruoyu Sun, and Yang Yu. Review of Reinforcement Learning for Large Language Models: Formulations, Algorithms, and Opportunities. *Transactions on Machine Learning Research*, 2026. URL https://www.liziniu.org/docs/RL4LLM_Survey.pdf.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. Self-distilled policy gradient, 2026. URL https://arxiv.org/abs/2606.04036.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023. URL https://arxiv.org/abs/2308.08747.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. WizardCoder: Empowering code large language models with Evol-Instruct, 2025. URL https://arxiv.org/abs/2306.08568.

Clayton Sanford, Bahare Fatemi, Ethan Hall, Anton Tsitsulin, Mehran Kazemi, Jonathan Halcrow, Bryan Perozzi, and Vahab Mirrokni. Understanding Transformer Reasoning Capabilities via Graph Algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8f395480c04ac6dfb2c2326a639df88e-Abstract-Conference.html.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. CRISP: Compressed reasoning via iterative self-policy distillation, 2026. URL https://arxiv.org/abs/2603.05433.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-Distillation Enables Continual Learning. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2601.19897.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's Razor: Why Online Reinforcement Learning Forgets Less. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/618c95f4557c15b253fb0e6f548ea0c0-Abstract-Conference.html.

Charles L. Wang, Keir Dorchen, and Peter Jin. On the statistical limits of self-improving agents, 2026a. URL https://arxiv.org/abs/2510.04399.

Hao Wang, Hao Gu, Hongming Piao, Kaixiong Gong, Yuxiao Ye, Xiangyu Yue, Sirui Han, Yike Guo, and Dapeng Wu. Entropy-preserving supervised fine-tuning via adaptive self-distillation for large reasoning models, 2026b. URL https://arxiv.org/abs/2602.02244.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal Subspace Learning for Language Model Continual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. TRACE: A Comprehensive Benchmark for Continual Learning in Large Language Models. *arXiv preprint arXiv:2310.06762*, 2023b. URL https://arxiv.org/abs/2310.06762.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023c. URL https://arxiv.org/abs/2203.11171.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning Large Language Models with Human: A Survey. *arXiv preprint arXiv:2307.12966*, 2023d. URL https://arxiv.org/abs/2307.12966.

Stanley Wei and Juno Kim. Provable Benefits of RLVR over SFT for Reasoning Models: Learning to Backtrack Efficiently. *arXiv preprint arXiv:2606.22938*, 2026. URL https://arxiv.org/abs/2606.22938. Preprint.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR, 2026. URL https://arxiv.org/abs/2604.03128.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 116194–116285. Curran Associates, Inc., 2025. doi: 10.52202/085713-3875. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a8f123480ede36612314d013aba6340c-Paper-Conference.pdf.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning, 2024. URL https://arxiv.org/abs/2402.13669.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning, 2025. URL https://arxiv.org/abs/2505.23433.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.10601.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026. URL https://arxiv.org/abs/2602.12275.

Huaye Zeng, Dongfu Jiang, Haozhe Wang, Ping Nie, Xiaotong Chen, and Wenhu Chen. ACE-CODER: Acing coder RL via automated test-case synthesis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12023–12040, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.587. URL https://aclanthology.org/2025.acl-long.587/.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction Tuning for Large Language Models: A Survey. *arXiv preprint arXiv:2308.10792*, 2023. URL https://arxiv.org/abs/2308.10792.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG, 2024. URL https://arxiv.org/abs/2403.10131.

Shiwan Zhao, Zhihu Wang, Xuyang Zhao, Jiaming Zhou, Caiyue Xu, Chenfei Liu, Liting Zhang, Yuhang Jia, Yanzhe Zhang, Hualong Yu, Zichen Xu, Qicheng Li, and Yong Qin. Large language model post-training: A unified view of off-policy and on-policy learning. *arXiv preprint arXiv:2604.07941*, 2026a. URL https://arxiv.org/abs/2604.07941.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2601.18734.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023. URL https://arxiv.org/abs/2303.18223.

Kejian Zhu, Zhuoran Jin, Shangqing Tu, Hongbang Yuan, Yushi Bai, Kang Liu, Juanzi Li, and Jun Zhao. SFT conflicts, RL coexists: A theoretical and empirical analysis of multi-task learning for LLMs. *arXiv preprint arXiv:2608.03573*, 2026. URL https://arxiv.org/abs/2608.03573.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.01347.

TABLE OF CONTENTS

1 Introduction **1**
1.1 Related work . . . 2

2 Problem Setup **3**
2.1 Continual learning on CoT graph . . . 3
2.2 Pre-training and fine-tuning . . . 4

3 Main results **5**
3.1 Single-task learnability . . . 6
3.2 Continual learning and catastrophic forgetting . . . 7
3.3 A theoretical implication for pre-training diversity . . . 8

4 Numerical experiments **9**

5 Conclusion **9**

Appendix A Additional related work **17**

Appendix B Realizability and proofs **17**
B.1 Realizability on routing-compatible layered DAGs . . . 18
Proposition B.1. Realizability on layered DAGs . . . **18**
Lemma 1. Positive routing with boundary coverage . . . 18
Lemma 2. Boundary kernels with small transition probabilities . . . 19
Lemma 3. Simultaneous finite softmax realization . . . 19
B.2 Gradient of the SD loss . . . 20
Lemma 4. Gradient of the SD loss . . . **20**
Lemma 5. The one-state stop-gradient identity . . . 20
Lemma 6. Cancellation of inactive hints . . . 21
B.3 Learning hint-supported inter-cluster transitions . . . 22
Proposition 2. Learning hint-supported transitions . . . **22**
Lemma 4. Gradient of the SD loss . . . 20
Lemma 7. Local updates and invariant interior transitions . . . 23
Lemma 8. Error from weighted update counts . . . 23
Lemma 9. Update opportunities from retained coverage . . . 24
Lemma 10. Simultaneous learning with finite mini-batches . . . 24
B.4 Single-task rollout accuracy under SD . . . 25
Theorem 1. Single-task learning under SD . . . **26**
Proposition 2. Learning hint-supported transitions . . . 22
Lemma 7. Local updates and invariant interior transitions . . . 23
Lemma 11. From edge accuracy to rollout accuracy . . . 25
B.5 Single-task learning under SFT . . . 26
Theorem 2. Single-task learning under SFT . . . **33**
Lemma 12. Sampled SFT gradient and update support . . . 27
Lemma 13. Convex potential, drift, and support error . . . 28
Lemma 14. A finite comparator . . . 29
Lemma 15. Conditional mini-batch noise . . . 30
Lemma 16. Smoothness and the initial transient . . . 31
Lemma 17. A finite-comparator last-iterate bound . . . 31
Lemma 18. Polynomial step-size sums . . . 32
Lemma 19. Simultaneous successor-set learning . . . 33
B.6 Retention under continual SD . . . 33
Theorem 3. No forgetting under SD . . . **34**
Proposition 2. Learning hint-supported transitions . . . 22
Lemma 11. From edge accuracy to rollout accuracy . . . 25
Lemma 20. Preservation across tasks . . . 34

B.7 Catastrophic forgetting under SFT . . . 34
Theorem 4. Catastrophic forgetting under SFT . . . **39**
Lemma 19. Simultaneous successor-set learning . . . 33
Lemma 21. Support separation and source overlap . . . 36
Lemma 22. Preservation of non-conflicted path mass . . . 37
Lemma 23. Exact transfer between task inputs . . . 37
Lemma 24. From successor suppression to old-path mass . . . 38
Remark 1. The old-path concentration profile . . . **39**
B.8 Pre-training and intermediate restartability . . . 40
Proposition 3. Uniform restartability . . . **40**
Lemma 10. Simultaneous learning with finite mini-batches . . . 24
Lemma 11. From edge accuracy to rollout accuracy . . . 25
Lemma 25. From training length to effective drift . . . 40
Proposition 1. Delayed restartability at a rare entrance . . . **42**
Lemma 9. Update opportunities from retained coverage . . . 24
Lemma 10. Simultaneous learning with finite mini-batches . . . 24
Lemma 11. From edge accuracy to rollout accuracy . . . 25
Lemma 25. From training length to effective drift . . . 40
Lemma 26. Control of visitation growth . . . 41
Lemma 27. Nonvisitation throughout adaptive training . . . 41
Lemma 28. An unvisited downstream cluster remains unlearned . . . 41
Remark 3. Separation of training times . . . **42**
Proposition 3. Uniform restartability . . . 40
Proposition 1. Delayed restartability at a rare entrance . . . 42
Remark 4. The scale of initial rarity . . . **43**
Lemma 29. The coarse rarity scale . . . 43
Remark 5. Correct paths bypassing the rare entrance . . . **43**
Proposition 1. Delayed restartability at a rare entrance . . . 42

Appendix C Numerical experiment **44**
C.1 Experiment settings . . . 44
C.2 Experiment results . . . 45

A Additional related work

In this section, we discuss additional related work.

On-policy versus off-policy learning. A central focus of this paragraph is the preservation of previously acquired capabilities. Continual SFT can degrade existing knowledge and reasoning abilities through catastrophic forgetting (Luo et al., 2023). Recent comparisons report that RL often preserves prior capabilities better than SFT at comparable new-task performance (Lai et al., 2025; Shenfeld et al., 2026b; Chen et al., 2025). Controlled experiments further identify on-policy sampling as an important contributor to this retention advantage (Chen et al., 2025). Zhu et al. (2026) attribute the contrasting forgetting behavior of SFT and RL to inter-task interference: SFT exhibits substantial conflicts during sequential training, whereas RL induces less interferences that allow multiple task capabilities to coexist. From a complementary distributional perspective, Shenfeld et al. (2026b) argue that standard SFT fits external demonstrations without explicitly accounting for proximity to the base model, whereas on-policy updates are biased toward solutions closer to the initial policy, helping explain their reduced forgetting. Building on Zhu et al. (2026)'s task-interference perspective, we develop more theoretical analysis of how SFT and on-policy SD differ in learning new tasks and preserving previously acquired capabilities.

CoT reasoning as graph search. Graph reasoning is a broadly accepted approach for analyzing the reasoning capabilities of language models. Sanford et al. (2024) characterize Transformer expressivity on graph reasoning tasks in terms of depth, width, and additional computation tokens, and empirically evaluate their theoretical predictions on GraphQA. Bachmann & Nagarajan (2024) use path-star graphs to demonstrate how teacher-forced next-token prediction can favor shortcuts that hinder learning to plan. Kim et al. (2025) model CoT as a metastable Markov process on clustered graphs and establish benefits of search, RL, and distillation for crossing rare inter-cluster transitions. Wei & Kim (2026) study backtracking in graph pathfinding and prove an exponential inference-time advantage of RLVR over SFT when SFT is trained only on golden shortest paths without negative examples. Our work uses a CoT graph model to analyze how SFT and self-distillation affect the retention of what they learned.

B Realizability and proofs

The proofs are organized as sketches, supporting lemmas, and main proofs. The algorithm of SD and SFT is described in the following.

Algorithm 1 Sequential SD and SFT in the reasoning-graph model

Require: method $\in \{\text{SD, SFT}\}$
Require: $\mathbf{W}^{(0)}$, B, $\{n_\ell\}_{\ell=1}^{L}$, $\{\eta_j\}_{j\geq 1}$, $\epsilon_{\text{rel}} \in (0,1)$
$\mathbf{W} \leftarrow \mathbf{W}^{(0)}$
for $\ell = 1, \ldots, L$ **do**
  for $j = 1, \ldots, n_\ell$ **do**
    for $b = 1, \ldots, B$ **do**
      Sample $y^{(b)} \sim \pi_{\mathbf{W}}(\cdot \mid x_\ell)$ and $\gamma^{(b)} \sim \mathrm{Unif}(\mathcal{P}_\ell)$.
      if method = SFT **then**
        $y^{*(b)} \leftarrow \gamma^{(b)}$
      else
        Construct $c^{(b)}$ from all inter-cluster transitions in $\gamma^{(b)}$.
        Construct $\pi_{\mathbf{W}}(\cdot \mid x_\ell, c^{(b)})$.
      end if
    end for
    $\mathbf{W} \leftarrow \mathbf{W} - \eta_j \nabla_{\mathbf{W}} L_{\text{method}}(\mathbf{W}; \mathcal{D}_\ell)$
  end for
end for
return $\mathbf{W}$

B.1 Realizability on routing-compatible layered DAGs

We consider layered DAGs satisfying the support conditions below. For this graph family, the proposition following the description shows that Assumption 2 can be realized by a finite pre-trained parameter $\mathbf{W}^{(0)}$. For $u \in V$, write $\mathcal{N}(u) := \{v \in V : (u, v) \in E\}$. Suppose that G is a finite layered DAG, meaning that

$$V = V_0 \dot\cup \cdots \dot\cup V_H, \qquad (u, v) \in E,\ u \in V_h,\ v \in V_{h'} \quad \Longrightarrow \quad h < h'. \tag{24}$$

We call such a graph *routing-compatible* for the task family if the following conditions hold:

1. Every boundary node has at least one same-cluster successor:

$$R_k(u) := \mathcal{N}(u) \cap C_k \neq \emptyset, \qquad u \in \partial C_k. \tag{25}$$

   These later-layer nodes provide support for the residual probability mass of a finite softmax distribution; they are not thereby declared relevant starts.

2. Every relevant start belongs to the interior of its cluster, and for every task $\mathcal{D}_\ell$, every non-terminal cluster C_k with $\partial^\star_\ell C_k \neq \emptyset$, and every relevant start s in C_k,

$$\partial^\star_\ell C_k \subseteq \mathcal{N}(s) \subseteq \partial C_k. \tag{26}$$

The numbers of layers, clusters, nodes, valid boundaries, and residual successors are arbitrary within this graph family.

Proposition B.1 (Realizability on routing-compatible layered DAGs)**.** *Suppose that G is a routing-compatible layered DAG in the sense of Equations 24–26. Under the full-rank condition on $(x_1, \ldots, x_L)$ in the problem setup, fix $\epsilon_1 \in (0, 1)$ and $\epsilon_2 > 0$ such that*

$$(K - 1)\epsilon_2 \leq \frac{1}{2}. \tag{27}$$

Then there exists a finite $\mathbf{W}^{(0)} \in \mathbb{R}^{d\times|E|}$ satisfying every clause of Assumption 2 simultaneously for all task inputs $x_1, \ldots, x_L$.

Sketch of proof of Proposition B.1.

- Construct positive within-cluster routing kernels with small invalid-boundary probability and uniform coverage of valid boundaries (Lemma 1).
- Construct normalized boundary kernels assigning the prescribed small probability to every inter-cluster edge (Lemma 2).
- Realize all task-wise kernels simultaneously with one finite softmax parameter (Lemma 3).

Lemma 1 (A positive routing kernel with boundary coverage)**.** *Under Equation 26, there is a strictly positive kernel at every relevant start for which the first boundary is reached in one step, the invalid-boundary probability is at most ϵ_1, and each valid boundary has probability at least $1/(2M)$.*

Proof. Fix a task $\mathcal{D}_\ell$, a non-terminal cluster C_k with $\partial^\star_\ell C_k \neq \emptyset$, and a relevant start s in C_k. Define

$$I_{\ell,k}(s) := \mathcal{N}(s) \setminus \partial^\star_\ell C_k. \tag{28}$$

By Equation 26, every outgoing neighbor of s is a boundary node and every valid boundary is an outgoing neighbor. Prescribe the transition kernel at s by

$$P_\ell(v \mid s) := \begin{cases} \dfrac{1}{(1+\epsilon_1)|\partial^\star_\ell C_k|}, & v \in \partial^\star_\ell C_k, \quad I_{\ell,k}(s) \neq \emptyset, \\ \dfrac{\epsilon_1}{(1+\epsilon_1)|I_{\ell,k}(s)|}, & v \in I_{\ell,k}(s), \quad I_{\ell,k}(s) \neq \emptyset, \\ \dfrac{1}{|\partial^\star_\ell C_k|}, & v \in \partial^\star_\ell C_k, \quad I_{\ell,k}(s) = \emptyset. \end{cases} \tag{29}$$

This is a strictly positive probability distribution on $\mathcal{N}(s)$. Because $\mathcal{N}(s) \subseteq \partial C_k$, the first boundary is reached in exactly one step, and hence

$$\mathbb{P}_{P_\ell}\big(U^s_{\ell,k} \notin \partial^\star_\ell C_k \mid Y_0 = s, x_\ell\big) \le \frac{\epsilon_1}{1+\epsilon_1} \le \epsilon_1, \tag{30}$$

$$\mathbb{P}_{P_\ell}\big(U^s_{\ell,k} = u \mid Y_0 = s, x_\ell\big) \ge \frac{1}{(1+\epsilon_1)|\partial^\star_\ell C_k|} \ge \frac{1}{2M} \tag{31}$$

for every $u \in \partial^\star_\ell C_k$. Thus the first-boundary hitting time is almost surely finite, the routing error is $O(\epsilon_1)$, and every valid boundary receives probability $\Omega(1/M)$, uniformly over all relevant starts. □

Lemma 2 (Boundary kernels with small inter-cluster probabilities)**.** *Under Equation 25 and Equation 27, each boundary admits a strictly positive normalized kernel assigning exactly ϵ_2 to every inter-cluster edge.*

Proof. We next prescribe the kernel at boundary nodes. For $u \in \partial C_k$, let

$$d^{\mathrm{out}}_{u,k} := |\mathcal{N}(u) \setminus C_k|. \tag{32}$$

For every task input, set

$$P_\ell(v \mid u) := \begin{cases} \epsilon_2, & v \in \mathcal{N}(u) \setminus C_k, \\ \dfrac{1 - d^{\mathrm{out}}_{u,k}\epsilon_2}{|R_k(u)|}, & v \in R_k(u). \end{cases} \tag{33}$$

The graph convention in the problem setup gives $d^{\mathrm{out}}_{u,k} \le K - 1$. Therefore Equation 27, together with Equation 25, makes Equation 33 a strictly positive probability distribution. □

Lemma 3 (Simultaneous finite softmax realization)**.** *Every family of strictly positive task-wise transition kernels P_ℓ on the given edge support is realized by one finite $\mathbf{W}^{(0)}$ when the task-input matrix has full column rank.*

Proof. For each task input, assign any strictly positive probability distribution to the outgoing edges of each remaining node whose kernel has not yet been specified. No kernel is needed at a terminal node with no outgoing edge. We now realize all these task-wise kernels with a single finite parameter matrix. Let

$$X := \begin{bmatrix} x_1^\top \\ \vdots \\ x_L^\top \end{bmatrix} \in \mathbb{R}^{L \times d}. \tag{34}$$

The full-rank input condition gives $\mathrm{rank}(X) = L$, so $XX^\top$ is invertible. For every $(u, v) \in E$, define

$$\ell_{u,v} := \begin{bmatrix} \log P_1(v \mid u) \\ \vdots \\ \log P_L(v \mid u) \end{bmatrix}, \qquad \mathbf{W}^{(0)} e_{u,v} := X^\top (XX^\top)^{-1} \ell_{u,v}. \tag{35}$$

All prescribed edge probabilities are strictly positive, so $\mathbf{W}^{(0)}$ is finite, and

$$x_\ell^\top \mathbf{W}^{(0)} e_{u,v} = \log P_\ell(v \mid u). \tag{36}$$

Since $P_\ell(\cdot \mid u)$ is normalized over $\mathcal{N}(u)$, the softmax predictor satisfies

$$\pi_{\mathbf{W}^{(0)}}(v \mid u, x_\ell) = \frac{\exp(\log P_\ell(v \mid u))}{\sum_{w \in \mathcal{N}(u)} \exp(\log P_\ell(w \mid u))} = P_\ell(v \mid u). \tag{37}$$

□

Main proof of Proposition B.1.

Proof. Lemmas 1 and 2 specify compatible kernels, since relevant starts are interior nodes. Lemma 3 realizes them simultaneously. Finally, Equation 24 implies that every transition strictly increases the layer index. Equations 30, 31, and 33 verify Assumption 2. □

Scope of the graph conditions. The support conditions above give a concrete finite-parameter realization of Assumption 2. However, our learning and forgetting results apply more broadly under their respective assumptions. Equation 25 ensures normalization when inter-cluster transitions are rare. Equation 26 simplifies the construction by ensuring that the first boundary is reached in one step. More generally, the within-cluster routing conditions allow multiple steps, provided boundary hitting is almost surely finite and the bounds on invalid-boundary probability and valid-boundary coverage in Equations 4 and 5 hold uniformly over relevant starts.

B.2 Gradient of the SD loss

Lemma 4 (Gradient of the SD loss)**.**

$$\nabla_{\mathbf{W}} L_{\mathrm{SD}}(\mathbf{W}; \mathcal{D}_\ell) = \frac{\log \epsilon_{\mathrm{rel}}}{B} \sum_{b=1}^{B} \sum_{i=1}^{I^{(b)}} \mathbf{1}\{\tau_i^{(b)} < T^{(b)}\} \pi_{\mathbf{W}}(v_i^{(b)} \mid u_i^{(b)}, x_\ell^{(b)}) \tag{38}$$

$$\cdot\, x_\ell^{(b)} \left(e_{u_i^{(b)}, v_i^{(b)}} - \sum_{v \in \mathcal{N}(u_i^{(b)})} \pi_{\mathbf{W}}(v \mid u_i^{(b)}, x_\ell^{(b)}) e_{u_i^{(b)}, v} \right)^\top \tag{39}$$

Sketch of proof of Lemma 4.

- Differentiate the one-state stop-gradient KL and show that successor-independent teacher factors contribute zero gradient (Lemma 5).
- Identify the active hint factor and show that the local gradient vanishes away from hint-source states (Lemma 6).
- Sum the local gradients over visited states and average over the mini-batch to obtain the stated sparse gradient formula (Lemma 4).

Lemma 5 (The one-state stop-gradient identity)**.** *For $p_t(v) = \pi_{\mathbf{W}}(v \mid y_t, x_\ell)$ and the stop-gradient teacher q_t, the KL and softmax derivatives satisfy Equation 42–Equation 43. In particular, a teacher factor independent of the proposed successor contributes zero gradient.*

Proof. For one batch element and one time $t < T$, set

$$u := y_t, \quad h_t := (y_0, \ldots, y_t), \quad p_t(v) := \pi_{\mathbf{W}}(v \mid u, x_\ell), \quad q_t(v) := \pi_{\mathbf{W}}(v \mid h_t, x_\ell, c).$$

With

$$\chi_{\mathbf{W}}(c \mid h_t, v, x_\ell) := \prod_{i=1}^{I} \chi_{\mathbf{W}}(c_i \mid h_t, v, x_\ell), \tag{40}$$

$$q_t(v) = \frac{\chi_{\mathbf{W}}(c \mid h_t, v, x_\ell) p_t(v)}{\sum_{z \in \mathcal{N}(u)} \chi_{\mathbf{W}}(c \mid h_t, z, x_\ell) p_t(z)}, \tag{41}$$

we have

$$\nabla_{\mathbf{W}} D_{\mathrm{KL}}(p_t \,\|\, \mathrm{SG}(q_t)) = \sum_{v \in \mathcal{N}(u)} \nabla_{\mathbf{W}} p_t(v) \log \frac{p_t(v)}{q_t(v)}, \tag{42}$$

$$\nabla_{\mathbf{W}} p_t(v) = p_t(v) x_\ell \left(e_{u,v} - \sum_{z \in \mathcal{N}(u)} p_t(z) e_{u,z} \right)^\top, \tag{43}$$

$$\sum_{v \in \mathcal{N}(u)} \nabla_{\mathbf{W}} p_t(v) = 0. \tag{44}$$

□

Lemma 6 (Cancellation of inactive hints)**.** *With the hint in Equation 9, the one-state gradient vanishes unless $t = \tau_i < T$ for a hint edge (u_i, v_i). At such a time it equals $\log(\epsilon_{\mathrm{rel}})\nabla_{\mathbf{W}}\pi_{\mathbf{W}}(v_i \mid u_i, x_\ell)$.*

Proof. By Lemma 5, it suffices to identify the successor-dependent hint factor. For each hint edge c_i,

$$\chi_{\mathbf{W}}(c_i \mid h_t, v, x_\ell) \propto \begin{cases} 1, & \tau_i = t,\ v = v_i, \\ \epsilon_{\mathrm{rel}}, & \tau_i = t,\ v \neq v_i, \\ 1, & \tau_i < t,\ y_{\tau_i+1} = v_i, \\ \epsilon_{\mathrm{rel}}, & \tau_i < t,\ y_{\tau_i+1} \neq v_i, \\ Z_{\mathbf{W},i}(x_\ell), & \tau_i > t, \end{cases}$$

where

$$Z_{\mathbf{W},i}(x_\ell) = \pi_{\mathbf{W}}(v_i \mid u_i, x_\ell) + \epsilon_{\mathrm{rel}}\left(1 - \pi_{\mathbf{W}}(v_i \mid u_i, x_\ell)\right).$$

Hence, if

$$t \notin \{\tau_1, \ldots, \tau_I\},$$

then

$$\chi_{\mathbf{W}}(c \mid h_t, v, x_\ell) \equiv \text{const. for } v \quad \Longrightarrow \quad q_t(v) = p_t(v) \quad \Longrightarrow \quad \nabla_{\mathbf{W}} D_{\mathrm{KL}}(p_t \,\|\, \mathrm{SG}(q_t)) = 0.$$

If $t = \tau_i < T$, node-simplicity of the sampled teacher path gives $\tau_{i'} \neq t$ for $i' \neq i$, and

$$Z_t := p_t(v_i) + \epsilon_{\mathrm{rel}}(1 - p_t(v_i)),$$
$$q_t(v) = \begin{cases} p_t(v)/Z_t, & v = v_i, \\ \epsilon_{\mathrm{rel}} p_t(v)/Z_t, & v \neq v_i. \end{cases}$$

Thus

$$\begin{aligned} \nabla_{\mathbf{W}} D_{\mathrm{KL}}(p_t \,\|\, \mathrm{SG}(q_t)) &= \sum_v \nabla_{\mathbf{W}} p_t(v) \log Z_t - (\log \epsilon_{\mathrm{rel}}) \sum_{v \neq v_i} \nabla_{\mathbf{W}} p_t(v) \\ &= (\log \epsilon_{\mathrm{rel}}) \nabla_{\mathbf{W}} p_t(v_i) \\ &= (\log \epsilon_{\mathrm{rel}}) p_t(v_i) x_\ell \left(e_{u_i, v_i} - \sum_{v \in \mathcal{N}(u_i)} p_t(v) e_{u_i, v} \right)^\top. \end{aligned}$$

□

Main proof of Lemma 4.

Proof. Apply Lemma 6 to each visited state. Node simplicity permits each hint source to contribute at most once. Averaging over the B paths gives

$$\begin{aligned} \nabla_{\mathbf{W}} L_{\mathrm{SD}}(\mathbf{W}; \mathcal{D}_\ell) = \frac{\log \epsilon_{\mathrm{rel}}}{B} \sum_{b=1}^{B} \sum_{i=1}^{I^{(b)}} \mathbf{1}\{\tau_i^{(b)} < T^{(b)}\} \pi_{\mathbf{W}}(v_i^{(b)} \mid u_i^{(b)}, x_\ell^{(b)}) \\ \cdot x_\ell^{(b)} \left(e_{u_i^{(b)}, v_i^{(b)}} - \sum_{v \in \mathcal{N}(u_i^{(b)})} \pi_{\mathbf{W}}(v \mid u_i^{(b)}, x_\ell^{(b)}) e_{u_i^{(b)}, v} \right)^\top. \end{aligned}$$

□

B.3 LEARNING HINT-SUPPORTED INTER-CLUSTER TRANSITIONS

We introduce the notation required for the following statement. Fix a task $\mathcal{D}_\ell$, and consider n consecutive SD updates on this task starting from $\mathbf{W}^{(0)}$. Let $N_\ell := |E_\ell|$, and enumerate the elements of E_ℓ as $e_m = (u_m, v_m)$, $m \in \{1, \dots, N_\ell\}$. Assumption 1 and the definition of M imply $N_\ell \le \sum_{k=1}^{K} |\partial_\ell^\star C_k| \le KM$.

Let $\mathcal{S}_{\ell,k} \subseteq \mathring{C}_k$ be the set of relevant starts in C_k: the task start $y_{\ell,0}$ when it belongs to C_k, an endpoint in C_k of a correct inter-cluster transition entering C_k. For $s \in \mathcal{S}_{\ell,k}$, write $U_{\ell,k}^s$ for the first boundary node reached by the rollout started from s, and define

$$\lambda_\ell := \min_{\substack{k:\, \partial_\ell^\star C_k \neq \emptyset \\ s \in \mathcal{S}_{\ell,k},\, u \in \partial_\ell^\star C_k}} \mathbb{P}_{\mathbf{W}^{(0)}}\left(U_{\ell,k}^s = u \mid s, x_\ell\right). \tag{45}$$

Assumption 2 implies the explicit, K-uniform bound $\lambda_\ell \ge c_{\mathrm{route}}/M > 0$. Lemma 4 shows that SD updates only edges whose sources are hint-source boundary nodes. Hence all transitions whose sources are interior nodes remain unchanged, and the first-boundary distribution in Equation 45 is preserved during SD training.

For each e_m, fix a shortest sequence $\mathcal{Q}_m = (C_{k_0}, \dots, C_{k_{L_m}})$ in the cluster-level graph induced by E_ℓ, from the cluster containing $y_{\ell,0}$ to the cluster containing u_m. The sequence can be chosen without repeated clusters, and therefore $L_m \le K - 1$. By Assumption 2, every correct inter-cluster edge has pretrained probability at least $c_{\mathrm{int}}\epsilon_2$. We therefore use the K-explicit lower bound

$$\underline{\rho}_{\ell,K,M} := \frac{1}{|\mathcal{P}_\ell|}\left(\frac{c_{\mathrm{route}}}{M}\right)^K (c_{\mathrm{int}}\epsilon_2)^{K-1} = \frac{c_{\mathrm{route}}^K c_{\mathrm{int}}^{K-1} \epsilon_2^{K-1}}{M^K |\mathcal{P}_\ell|}. \tag{46}$$

Under Equation 6, every path in $\mathcal{P}_\ell$ is sampled with probability exactly $1/|\mathcal{P}_\ell|$. The conditional independence properties used below are exactly those specified in the fine-tuning setup above. Let $\{\eta_j\}_{j\ge 1}$ be a diminishing positive step-size sequence, and let $\delta \in (0,1)$. Define the finite-sample effective drift by

$$\mathfrak{D}_{\ell,n}^{(K)}(\delta) := \left[\underline{\rho}_{\ell,K,M} \sum_{j=1}^{n} \eta_j - \sqrt{\frac{\log(KM/\delta)}{2B} \sum_{j=1}^{n} \eta_j^2}\right]^+. \tag{47}$$

Proposition 2 (Learning hint-supported inter-cluster transitions)**.** *Suppose Assumptions 1 and 2 hold and $0 < \epsilon_2 \le 1$. With probability at least $1 - \delta$, simultaneously for every $(u_m, v_m) \in E_\ell$,*

$$\pi_{\mathbf{W}^{(n)}}(v_m \mid u_m, x_\ell) \ge 1 - O\left(\left[\epsilon_2 + \epsilon_2^2 \left(\log \frac{1}{\epsilon_{\mathrm{rel}}}\right) \mathfrak{D}_{\ell,n}^{(K)}(\delta)\right]^{-1}\right), \tag{48}$$

where $[a]_+ := \max\{a, 0\}$. If in addition $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in [0, 1)$, then for sufficiently large n the drift term dominates the fluctuation term and

$$\pi_{\mathbf{W}^{(n)}}(v_m \mid u_m, x_\ell) \ge 1 - O\left(\left[\epsilon_2 + \epsilon_2^2 \left(\log \frac{1}{\epsilon_{\mathrm{rel}}}\right) \underline{\rho}_{\ell,K,M} n^{1-\nu}\right]^{-1}\right). \tag{49}$$

The learning term inside the brackets has the explicit M, K dependence

$$\epsilon_2^2 \underline{\rho}_{\ell,K,M} n^{1-\nu} = \frac{c_{\mathrm{route}}^K c_{\mathrm{int}}^{K-1}}{M^K |\mathcal{P}_\ell|} \epsilon_2^{K+1} n^{1-\nu}.$$

Sketch of proof of Proposition 2.

- Show that local SD updates increase correct logit gaps and preserve the interior transition kernel (Lemma 7).
- Use retained boundary coverage to lower-bound the probability of an update opportunity for each correct edge (Lemma 9).
- Control the weighted update counts simultaneously using conditional mini-batch concentration (Lemma 10).
- Convert the weighted update counts into transition-error bounds through the odds estimate (Lemma 8).

Shared notation and constants. The local update and odds estimates below apply under either Assumptions 1 and 2 or Assumptions 1 and 3. Unless stated otherwise, the SD lemmas concern training on the fixed task alone. Throughout the SD lemmas, $C_{\mathrm{bad}}\epsilon_1$ bounds the initial invalid-boundary probability and $c_{\mathrm{int}}\epsilon_2$ bounds every initial correct inter-cluster probability from below. The constants are positive and uniform; we may take $c_{\mathrm{int}} \le 1$ and $c_{\mathrm{route}}/M \le 1$. For the rare-entrance case, let $e_m = (u_m, v_m)$ be the edge in Assumption 3, so $u_m = u^*$. The first-boundary lower bound is retained at all other correct boundaries; c_{route} and $\underline{\rho}_{\ell,K,M}$ use only this retained lower bound in the rare case.

Let $\mathcal{F}_{j-1}$ be the training history before iteration j. Define

$$\tau_i^{(b,j)} = \inf\{t \ge 0 : y_t^{(b,j)} = u_i\}, \qquad I_i^{(b,j)} = \mathbf{1}\{e_i \in c^{(b,j)}\}\mathbf{1}\{\tau_i^{(b,j)} < T^{(b,j)}\},$$
$$\mathfrak{n}_i^{(j)} = \sum_{b=1}^{B} I_i^{(b,j)}, \qquad p_i^{(j)} = p_{u_i,v_i}^{(j)},$$
$$p_{u,v}^{(j)} = \pi_{\mathbf{W}^{(j)}}(v \mid u, x_\ell), \qquad z_{u,v}^{(j)} = x_\ell^\top \mathbf{W}^{(j)} e_{u,v},$$
$$r_i^{(j)} = (1 - p_i^{(j)})/p_i^{(j)}, \qquad a_{i,j} = \frac{\eta_j}{B}\log(1/\epsilon_{\mathrm{rel}})\mathfrak{n}_i^{(j)} p_i^{(j-1)}.$$

The symbols $e_{u,v}$ denote one-hot edge features; e_i denotes a graph edge. Assumption 1 gives distinct sources u_i and $N_\ell = \sum_h |\partial_\ell^\star C_h| \le KM$.

Lemma 7 (Local update and invariant interior transitions)**.** *For every training history and every outgoing edge of u_i,*

$$\mathbf{W}^{(j)} e_{u_i,v} = \mathbf{W}^{(j-1)} e_{u_i,v} + a_{i,j}(\mathbf{1}\{v = v_i\} - p_{u_i,v}^{(j-1)})x_\ell. \tag{50}$$

Only columns whose sources occur in E_ℓ can change. Moreover,

$$p_i^{(j)} \ge p_i^{(j-1)}, \qquad \frac{p_{u_i,v}^{(j)}}{p_i^{(j)}} \le \frac{p_{u_i,v}^{(j-1)}}{p_i^{(j-1)}}. \tag{51}$$

First-boundary hitting distributions from relevant starts remain equal to their initial values.

Proof. Lemma 4 gives the visited-source contribution

$$(\log \epsilon_{\mathrm{rel}}) p_i^{(j-1)} x_\ell \left(e_{u_i,v_i} - \sum_{v \in \mathcal{N}(u_i)} p_{u_i,v}^{(j-1)} e_{u_i,v} \right)^\top.$$

The hint's normalization term vanishes because $\sum_v \nabla p_{u_i,v} = 0$. At other sources the hinted and unhinted distributions agree, giving zero gradient. Uniqueness, node simplicity, and mini-batch averaging yield Equation 50. For $v \ne v_i$, the correct-minus-incorrect logit gap increases by

$$a_{i,j}\|x_\ell\|_2^2(1 - p_i^{(j-1)} + p_{u_i,v}^{(j-1)}) \ge 0.$$

Exponentiating proves the ratio inequality, and summing the incorrect-to-correct ratios proves monotonicity of $p_i^{(j)}$. Interior sources never occur in hints, so all transitions before the first boundary are unchanged. This also preserves boundary-hitting properness. □

Lemma 8 (Error from weighted update counts)**.** *For every history and every i,*

$$1 - p_i^{(n)} \le \left[1 + c_{\mathrm{int}}^2 \epsilon_2^2 \log(1/\epsilon_{\mathrm{rel}}) \frac{1}{B} \sum_{j=1}^{n} \eta_j \mathfrak{n}_i^{(j)} \right]^{-1}. \tag{52}$$

If $p_i^{(0)} < 1$, the stronger estimate is

$$r_i^{(n)} \le \left[(r_i^{(0)})^{-1} + \|x_\ell\|_2^2 (p_i^{(0)})^2 \log(1/\epsilon_{\mathrm{rel}}) \frac{1}{B} \sum_{j=1}^{n} \eta_j \mathfrak{n}_i^{(j)} \right]^{-1}. \tag{53}$$

Proof. If $p_i^{(0)} = 1$, the assertion is immediate. Otherwise the logit-gap estimate above gives $r_i^{(j)} \le r_i^{(j-1)} \exp[-a_{i,j}\|x_\ell\|_2^2(1-p_i^{(j-1)})]$. Since $r_i^{(j-1)} \le r_i^{(0)}$ and $p_i^{(j-1)}(1-p_i^{(j-1)}) \ge r_i^{(j-1)}/(1+r_i^{(0)})^2$, the inequality $e^t \ge 1+t$ implies

$$\frac{1}{r_i^{(j)}} \ge \frac{1}{r_i^{(j-1)}} + \frac{\eta_j}{B}\|x_\ell\|_2^2 (p_i^{(0)})^2 \log(1/\epsilon_{\mathrm{rel}})\mathfrak{n}_i^{(j)}.$$

Telescoping proves Equation 53. For the error bound, use $p_i^{(0)} \ge c_{\mathrm{int}}\epsilon_2$ and $\|x_\ell\|_2^2 \ge 1$, and write $1 - p_i^{(n)} = (1 + 1/r_i^{(n)})^{-1}$. □

Lemma 9 (Update opportunities from retained coverage)**.** *Under Assumptions 1 and 2, every $e_i \in E_\ell$ satisfies*

$$\mathbb{E}[I_i^{(b,j)} \mid \mathcal{F}_{j-1}] \ge \underline{\rho}_{\ell,K,M}. \tag{54}$$

Under Assumptions 1 and 3, the same holds for every $e_i \ne e_m$ whose source cluster is reachable from the original-start cluster in the cluster graph induced by $E_\ell \setminus \{e_m\}$.

Proof. Fix an edge in the specified set. Choose a shortest sequence of clusters leading to its source cluster, using correct inter-cluster edges (excluding e_m in the second case). The sequence has no repeated cluster and at most $K-1$ transitions. Specify one such edge between each adjacent pair, and finally specify boundary u_i in the last cluster. In the second case all specified boundary sources differ from u_m, since its unique correct edge is e_m. Also, the sequence cannot enter C_k by Assumption 3.

Every internal segment begins at a relevant start. Its probability of first reaching the specified boundary is at least c_{route}/M, by retained coverage and Lemma 7. At each required crossing the correct edge probability is at least $c_{\mathrm{int}}\epsilon_2$, by monotonicity. Apply the Markov property successively at these boundary and entry times. There are at most K internal segments and $K-1$ crossings, so

$$\mathbb{P}(\tau_i^{(b,j)} < T^{(b,j)} \mid \mathcal{F}_{j-1}) \ge (c_{\mathrm{route}}/M)^K (c_{\mathrm{int}}\epsilon_2)^{K-1}.$$

Both lower-bound factors can be taken at most one. Since $e_i \in E_\ell$, at least one correct path contains e_i, so the uniform hint contains it with probability at least $1/|\mathcal{P}_\ell|$. The stipulated conditional independence of hint and student gives Equation 54. □

Lemma 10 (Simultaneous learning with finite mini-batches)**.** *For the applicable edge set in Lemma 9, with probability at least $1-\delta$, every such edge satisfies*

$$\frac{1}{B}\sum_{j=1}^{n} \eta_j \mathfrak{n}_i^{(j)} \ge \mathfrak{D}_{\ell,n}^{(K)}(\delta). \tag{55}$$

Combining this event with Lemma 8 gives simultaneous error bounds for all these edges.

Proof. For a fixed i, conditional mini-batch independence and Hoeffding's lemma give, for $t > 0$,

$$\mathbb{E}\left[\exp\left\{-\frac{t\eta_j}{B}\sum_{b=1}^{B}(I_i^{(b,j)} - \mathbb{E}[I_i^{(b,j)} \mid \mathcal{F}_{j-1}])\right\}\middle| \mathcal{F}_{j-1}\right] \le \exp\left(\frac{t^2\eta_j^2}{8B}\right).$$

Iterating this bound over j gives the exponent $t^2\sum_j \eta_j^2/(8B)$ for the full sum. Exponential Markov and optimization over t imply

$$\mathbb{P}\left(\frac{1}{B}\sum_{j=1}^{n}\eta_j\sum_{b=1}^{B}[I_i^{(b,j)} - \mathbb{E}[I_i^{(b,j)} \mid \mathcal{F}_{j-1}]] < -\sqrt{\frac{\log(KM/\delta)}{2B}\sum_{j=1}^{n}\eta_j^2}\right) \le \frac{\delta}{KM}.$$

Use Lemma 9 and union-bound over at most KM edges. Nonnegativity of the weighted counts permits taking the positive part, giving Equation 55. Lemma 8 gives the error bound. If the edge set is empty, the simultaneous assertion is vacuous. □

Lemma 11 (Converting edge accuracy to rollout accuracy)**.** *Under Assumptions 1 and 2, consider any predictor whose interior kernel equals the pretrained kernel, and denote its edge probabilities by $p_i^{(n)}$. Then every relevant start $s \in \mathcal{S}_{\ell,h}$ satisfies*

$$1 - \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell) \le C_{\mathrm{bad}}(1 + N_\ell)\epsilon_1 + \sum_{i=1}^{N_\ell}(1 - p_i^{(n)}). \tag{56}$$

Under Assumptions 1 and 3, if the edges specified in Lemma 9 have error at most ϵ_1, then

$$\begin{aligned} &1 - \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) \\ &\quad \le C_{\mathrm{bad}}(1 + KM)\epsilon_1 + KM\epsilon_1 + \pi_{\mathbf{W}^{(n)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell). \end{aligned} \tag{57}$$

The first inequality is deterministic and does not require the predictor to have been trained on the fixed task alone.

Proof. For Equation 56, follow a rollout until reaching the target or the first invalid boundary or incorrect choice at a valid boundary. Before these events, each internal segment starts at a relevant start and is followed by a correct inter-cluster edge. Conditional on the history up to each segment's start, the invalid-boundary probability is at most $C_{\mathrm{bad}}\epsilon_1$. By node simplicity there are at most $1+N_\ell$ segments and each correct source is visited at most once. Summing conditional probabilities, including the events that the segments actually start, gives respectively $C_{\mathrm{bad}}(1+N_\ell)\epsilon_1$ and $\sum_i(1 - p_i^{(n)})$. An incorrect choice includes choosing an intra-cluster edge at a valid boundary.

These events cover failure: any reached nonterminal cluster has a correct exit, because the arriving correct edge lies on a correct path with a remaining suffix. The analogous statement holds for the original start. Properness makes each internal segment finite. If the target were never reached while both errors were avoided, one would traverse more than N_ℓ distinct correct sources, contradicting node simplicity. For a start already in the target the conclusion is immediate.

For Equation 57, start from $y_{\ell,0}$ and stop additionally upon visiting u_m. Before that visit every current cluster is reachable via $E_\ell \setminus \{e_m\}$, so each encountered correct boundary other than u_m belongs to the learned set in Lemma 9. Its choice error is at most ϵ_1. The invalid-boundary term is unchanged, and the extra stopping event is bounded by the unconditional evaluation visitation probability displayed in Equation 57. The same properness argument covers failure. □

Main proof of Proposition 2.

Proof. Lemmas 9 and 10 give $B^{-1}\sum_j \eta_j \mathfrak{n}_i^{(j)} \ge \mathfrak{D}_{\ell,n}^{(K)}(\delta)$ simultaneously for all i, with probability at least $1-\delta$. If $p_i^{(0)} = 1$, the transition is already correct with probability one. Otherwise, $(r_i^{(0)})^{-1} = p_i^{(0)}/(1-p_i^{(0)}) \ge c_{\mathrm{int}}\epsilon_2$ and $(p_i^{(0)})^2 \ge c_{\mathrm{int}}^2\epsilon_2^2$. The stronger estimate in Lemma 8, together with $1 - p_i^{(n)} \le r_i^{(n)}$, proves Equation 48. For fixed $0 \le \nu < 1$, $\sum_{j=1}^n \eta_j = \Theta(n^{1-\nu})$, whereas $(\sum_{j=1}^n \eta_j^2)^{1/2}$ is respectively $O(n^{1/2-\nu})$, $O(\sqrt{\log(en)})$, or $O(1)$ as $\nu < 1/2$, $\nu = 1/2$, or $\nu > 1/2$. Thus the fluctuation is smaller than the drift for sufficiently large n, which proves Equation 49. □

B.4 Single-task rollout accuracy under SD

Theorem 1 (Single-task learning under SD)**.** *Fix a task $\mathcal{D}_\ell$ and let the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in [0, 1/2)$. Suppose Assumptions 1 and 2 hold and ϵ_1 is sufficiently small. Define*

$$\xi_{\ell,n}^{(K)}(\delta) := \left[\epsilon_2 + \epsilon_2^2\left(\log\frac{1}{\epsilon_{\mathrm{rel}}}\right)\left[\frac{e^{-\Theta(K\log\frac{1}{\epsilon_2})}n^{1-\nu}}{M^K|\mathcal{P}_\ell|} - \sqrt{\frac{\log(KM/\delta)}{B}}n^{1/2-\nu}\right]^+\right]^{-1}. \tag{11}$$

Then, with probability at least $1-\delta$,

$$\pi_{\mathbf{W}^{(n)}}\left(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell\right) \ge 1 - O(KM\epsilon_1) - O\Big(KM\xi_{\ell,n}^{(K)}(\delta)\Big). \tag{12}$$

Sketch of proof of Theorem 1.

- Bound the errors of all correct inter-cluster transitions on one common training event (Proposition 2).
- Show that the interior transition kernel retains its pre-trained routing guarantees (Lemma 7).
- Bound rollout failure by the probabilities of invalid-boundary encounters and incorrect transition choices (Lemma 11).

Main proof of Theorem 1.

Proof. Proposition 2 gives an event of probability at least $1-\delta$ on which $1-p_i^{(n)} \le C_{\mathrm{tr}}\xi_{\ell,n}^{(K)}(\delta)$ for every i. Lemma 7 preserves the interior kernel. Apply Lemma 11 with $s = y_{\ell,0}$ to obtain

$$\begin{aligned} 1-\pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) &\le C_{\mathrm{bad}}(1+N_\ell)\epsilon_1 + N_\ell C_{\mathrm{tr}}\xi_{\ell,n}^{(K)}(\delta) \\ &= O(KM\epsilon_1) + O(KM\xi_{\ell,n}^{(K)}(\delta)). \end{aligned}$$

This proves Equation 12 on that same event. In the regime where the learning term dominates, substitution of Equation 46 gives

$$KM\xi_{\ell,n}^{(K)}(\delta) = O\left(\frac{KM^{K+1}|\mathcal{P}_\ell|}{c_{\mathrm{route}}^K c_{\mathrm{int}}^{K-1}\epsilon_2^{K+1} n^{1-\nu}\log(1/\epsilon_{\mathrm{rel}})}\right).$$

□

B.5 Single-task learning under SFT

Set

$$m_\nu := \min\{\nu, 1-\nu\}.$$

We prove the stronger batch-explicit bound

$$\min_{u\in\mathcal{U}_\ell} \pi_{\mathbf{W}_{\mathrm{SFT}}^{(n)}}(\mathcal{A}_\ell(u) \mid u, x_\ell) \ge 1 - C_\ell|\mathcal{P}_\ell|\left[\frac{1}{n} + \frac{\log^2(e+n|\mathcal{P}_\ell|)}{n^{1-\nu}} + \frac{(1+2\log(2/\delta))\log(en)}{B\, n^{m_\nu}}\right], \tag{58}$$

where C_ℓ is independent of n, B, δ.

Let Γ be one teacher path. Define

$$\omega_{\ell,u} := \mathbb{E}[N_u(\Gamma)], \quad r_{\ell,u}(v) := \frac{\mathbb{E}[N_{u,v}(\Gamma)]}{\omega_{\ell,u}}, \qquad v \in \mathcal{A}_\ell(u). \tag{59}$$

Then

$$\omega_{\ell,u} \ge \frac{1}{|\mathcal{P}_\ell|}, \quad r_{\ell,u}(v) > 0. \tag{60}$$

Theorem 2 (Single-task learning under SFT)**.** *Fix $\delta \in (0,1)$. Suppose the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in (0, 1/2)$ and $\mathbf{W}_{\mathrm{SFT}}^{(0)} = \mathbf{W}^{(0)}$. Then, with probability at least $1-\delta$,*

$$\min_{u\in\mathcal{U}_\ell} \pi_{\mathbf{W}_{\mathrm{SFT}}^{(n)}}(\mathcal{A}_\ell(u) \mid u, x_\ell) \ge 1 - \tilde{O}\left(|\mathcal{P}_\ell|\left[\frac{1}{n^{1-\nu}} + \frac{\log(e/\delta)}{Bn^\nu}\right]\right). \tag{13}$$

Here $\mathcal{U}_\ell$ is the set of nodes appearing in the correct paths except for the terminal, and $\mathcal{A}_\ell(u)$ is the set of their correct successors from u. $\tilde{O}(\cdot)$ hides logarithmic factors in n and $|\mathcal{P}_\ell|$.

Sketch of proof of Theorem 2.

- Derive the sampled SFT gradient and identify a convex potential whose gradient is its conditional expectation (Lemmas 12 and 13).
- Construct a finite comparator with small potential and a controlled distance from the initial parameter (Lemma 14).
- Verify the conditional mini-batch noise and smoothness bounds, including the bounded initial transient (Lemmas 15 and 16).
- Apply the finite-comparator last-iterate bound and evaluate the polynomial step-size sums (Lemmas 17 and 18).
- Convert the resulting potential bound into successor-set accuracy simultaneously at every teacher source (Lemmas 13 and 19).

At iteration j,

$$L_{\mathrm{SFT}}^{(j)}(\mathbf{W}) := -\frac{1}{B}\sum_{b=1}^{B}\sum_{t=0}^{T^{(b,j)}-1}\log \pi_{\mathbf{W}}(y_{t+1}^{*(b,j)} \mid y_t^{*(b,j)}, x_\ell), \tag{61}$$

$$\mathbf{g}_j := \nabla_{\mathbf{W}} L_{\mathrm{SFT}}^{(j)}(\mathbf{W}_{\mathrm{SFT}}^{(j-1)}),$$

$$\mathbf{W}_{\mathrm{SFT}}^{(j)} := \mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \eta_j \mathbf{g}_j. \tag{62}$$

Lemma 12 (Exact sampled SFT gradient and update support)**.** *The mini-batch update Equation 62 has gradient Equation 63, whose Frobenius norm is at most $G_\ell = \sqrt{2d(|V|-1)}$. Every update lies in $\mathrm{span}\{x_\ell\}$ within each source block, and only sources in $\mathcal{U}_\ell$ can change.*

Proof. Direct differentiation gives

$$\mathbf{g}_j = \frac{1}{B}\sum_{b=1}^{B}\sum_{u\in\mathcal{U}_\ell} x_\ell \left(N_u(\Gamma_j^{(b)}) \sum_{v\in\mathcal{N}(u)} \pi_{\mathbf{W}_{\mathrm{SFT}}^{(j-1)}}(v \mid u, x_\ell) e_{u,v} - \sum_{v\in\mathcal{A}_\ell(u)} N_{u,v}(\Gamma_j^{(b)}) e_{u,v} \right)^\top. \tag{63}$$

For one path Γ, let $\mathbf{h}(\Gamma;\mathbf{W})$ denote the corresponding single-path gradient. If $v_\Gamma(u)$ is the successor of the visited source u, then

$$\left\| \sum_{v\in\mathcal{N}(u)} \pi_{\mathbf{W}}(v \mid u, x_\ell) e_{u,v} - e_{u,v_\Gamma(u)} \right\|_2^2 = \sum_{v\in\mathcal{N}(u)} \pi_{\mathbf{W}}(v \mid u, x_\ell)^2 + 1 - 2\pi_{\mathbf{W}}(v_\Gamma(u) \mid u, x_\ell)$$
$$\leq 2.$$

Since the source-node edge blocks are disjoint,

$$\|\mathbf{h}(\Gamma;\mathbf{W})\|_{\mathrm{F}}^2 \leq 2\|x_\ell\|_2^2(|\Gamma|-1) \leq 2d(|V|-1).$$

Hence

$$\|\mathbf{g}_j\|_{\mathrm{F}} \leq G_\ell = \sqrt{2d(|V|-1)}. \tag{64}$$

□

Define

$$P_{\ell,u}(\mathbf{W}) := \pi_{\mathbf{W}}(\mathcal{A}_\ell(u) \mid u, x_\ell),$$

and

$$F_\ell(\mathbf{W}) := \sum_{u\in\mathcal{U}_\ell} \omega_{\ell,u} \left[D_{\mathrm{KL}}\left(r_{\ell,u} \middle\| \frac{\pi_{\mathbf{W}}(\cdot \mid u, x_\ell)}{P_{\ell,u}(\mathbf{W})} \right) - \log P_{\ell,u}(\mathbf{W}) \right]. \tag{65}$$

Lemma 13 (Convex potential, unbiased drift, and support error)**.** *The potential F_ℓ in Equation 65 is nonnegative and convex, the sampled gradient satisfies Equation 67, and* $\max_{u\in\mathcal{U}_\ell}[1-P_{\ell,u}(\mathbf{W})] \le |\mathcal{P}_\ell|F_\ell(\mathbf{W})$ *for every finite* $\mathbf{W}$.

Proof. For every u,

$$D_{\mathrm{KL}}\left(r_{\ell,u}\middle\|\frac{\pi_{\mathbf{W}}}{P_{\ell,u}}\right) - \log P_{\ell,u}$$
$$= \sum_{v\in\mathcal{A}_\ell(u)} r_{\ell,u}(v)\log r_{\ell,u}(v) - \sum_{v\in\mathcal{A}_\ell(u)} r_{\ell,u}(v)\log \pi_{\mathbf{W}}(v \mid u, x_\ell).$$

Therefore

$$F_\ell(\mathbf{W}) = C'_\ell - \sum_{u\in\mathcal{U}_\ell}\sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)]\log \pi_{\mathbf{W}}(v \mid u, x_\ell), \tag{66}$$

so F_ℓ is convex. We next verify the conditional-drift identity explicitly. Let $\mathcal{F}_{j-1}$ be the sigma-field generated by all training randomness up to iteration $j-1$. Conditional on $\mathcal{F}_{j-1}$, the parameter $\mathbf{W}_{\mathrm{SFT}}^{(j-1)}$ is fixed, while $\Gamma_j^{(1)},\dots,\Gamma_j^{(B)}$ are conditionally independent and each is uniformly distributed on $\mathcal{P}_\ell$. Therefore, for every $u\in\mathcal{U}_\ell$, $v\in\mathcal{A}_\ell(u)$, and $b\in\{1,\dots,B\}$,

$$\mathbb{E}\Big[N_u(\Gamma_j^{(b)}) \mid \mathcal{F}_{j-1}\Big] = \mathbb{E}[N_u(\Gamma)] = \omega_{\ell,u}, \qquad \text{by conditional uniform sampling,}$$
$$\mathbb{E}\Big[N_{u,v}(\Gamma_j^{(b)}) \mid \mathcal{F}_{j-1}\Big] = \mathbb{E}[N_{u,v}(\Gamma)] = \omega_{\ell,u} r_{\ell,u}(v), \qquad \text{by Equation 59.}$$

Moreover, for every correct teacher path Γ,

$$N_u(\Gamma) = \sum_{v\in\mathcal{A}_\ell(u)} N_{u,v}(\Gamma), \qquad \text{each source occurrence has one correct successor.}$$

Taking expectations gives

$$\omega_{\ell,u} = \sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)].$$

Now apply conditional expectation to Equation 63. Write $\mathbf{W}_{j-1} := \mathbf{W}_{\mathrm{SFT}}^{(j-1)}$ only within the following calculation. Since $\mathbf{W}_{j-1}$ is $\mathcal{F}_{j-1}$-measurable, the softmax probabilities are fixed under the conditional expectation. Substituting the two conditional-expectation identities above into Equation 63 gives

$$\mathbb{E}[\mathbf{g}_j \mid \mathcal{F}_{j-1}] = \frac{1}{B}\sum_{b=1}^{B}\sum_{u\in\mathcal{U}_\ell} x_\ell \left(\omega_{\ell,u}\sum_{z\in\mathcal{N}(u)} \pi_{\mathbf{W}_{j-1}}(z \mid u, x_\ell)e_{u,z} - \sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)]e_{u,v}\right)^\top$$
$$= \sum_{u\in\mathcal{U}_\ell} x_\ell \left(\omega_{\ell,u}\sum_{z\in\mathcal{N}(u)} \pi_{\mathbf{W}_{j-1}}(z \mid u, x_\ell)e_{u,z} - \sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)]e_{u,v}\right)^\top.$$

The second equality holds because the summand no longer depends on b, so the factor B^{-1} cancels with the sum over the B mini-batch elements.

We next differentiate the potential. Equation 66 gives

$$\nabla F_\ell(\mathbf{W}) = -\sum_{u\in\mathcal{U}_\ell}\sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)]\nabla_{\mathbf{W}}\log \pi_{\mathbf{W}}(v \mid u, x_\ell),$$

because C'_ℓ is independent of $\mathbf{W}$. For the softmax predictor,

$$-\nabla_{\mathbf{W}}\log \pi_{\mathbf{W}}(v \mid u, x_\ell) = x_\ell\left(\sum_{z\in\mathcal{N}(u)} \pi_{\mathbf{W}}(z \mid u, x_\ell)e_{u,z} - e_{u,v}\right)^\top.$$

Substituting this identity and collecting the terms with the same source u yields

$$\nabla F_\ell(\mathbf{W}) = \sum_{u\in\mathcal{U}_\ell} x_\ell \Bigg(\omega_{\ell,u} \sum_{z\in\mathcal{N}(u)} \pi_{\mathbf{W}}(z \mid u, x_\ell) e_{u,z} - \sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)] e_{u,v} \Bigg)^\top.$$

Here we used $\omega_{\ell,u} = \sum_{v\in\mathcal{A}_\ell(u)} \mathbb{E}[N_{u,v}(\Gamma)]$, which follows from the preceding path-count identity. Evaluating this identity at $\mathbf{W} = \mathbf{W}_{\mathrm{SFT}}^{(j-1)}$ proves

$$\mathbb{E}[\mathbf{g}_j \mid \mathcal{F}_{j-1}] = \nabla F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(j-1)}). \tag{67}$$

Using

$$D_{\mathrm{KL}}(\cdot\|\cdot) \geq 0, \quad -\log z \geq 1 - z,$$

we obtain, for every u,

$$F_\ell(\mathbf{W}) \geq \omega_{\ell,u}[-\log P_{\ell,u}(\mathbf{W})] \geq \frac{1}{|\mathcal{P}_\ell|}[1 - P_{\ell,u}(\mathbf{W})].$$

Thus

$$\max_{u\in\mathcal{U}_\ell}[1 - P_{\ell,u}(\mathbf{W})] \leq |\mathcal{P}_\ell| F_\ell(\mathbf{W}). \tag{68}$$

□

Lemma 14 (A finite comparator approaching the optimal support)**.** *For every $R > 0$, there is a finite $\mathbf{W}_R$ satisfying $F_\ell(\mathbf{W}_R) \leq C_0 e^{-R}$ and $\|\mathbf{W}_R - \mathbf{W}^{(0)}\|_{\mathrm{F}}^2 \leq C_1(1 + R + \log|\mathcal{P}_\ell|)^2$, with constants independent of R, n, B, δ.*

Proof. For $R > 0$, set

$$z_{u,v}^{(R)} = \begin{cases} R + \log r_{\ell,u}(v), & v \in \mathcal{A}_\ell(u), \\ 0, & v \notin \mathcal{A}_\ell(u). \end{cases} \tag{69}$$

Let

$$m_{\ell,u}^- := |\mathcal{N}(u) \setminus \mathcal{A}_\ell(u)|.$$

Then

$$\sum_{v\in\mathcal{A}_\ell(u)} e^{z_{u,v}^{(R)}} = e^R,$$

$$\sum_{v\notin\mathcal{A}_\ell(u)} e^{z_{u,v}^{(R)}} = m_{\ell,u}^-,$$

$$P_{\ell,u}(\mathbf{W}_R) = \frac{e^R}{e^R + m_{\ell,u}^-} = \frac{1}{1 + m_{\ell,u}^- e^{-R}},$$

$$\frac{\pi_{\mathbf{W}_R}(v \mid u, x_\ell)}{P_{\ell,u}(\mathbf{W}_R)} = r_{\ell,u}(v), \qquad v \in \mathcal{A}_\ell(u).$$

Hence

$$\begin{aligned} F_\ell(\mathbf{W}_R) &= \sum_{u\in\mathcal{U}_\ell} \omega_{\ell,u} \log(1 + m_{\ell,u}^- e^{-R}) \\ &\leq e^{-R} \sum_{u\in\mathcal{U}_\ell} \omega_{\ell,u} m_{\ell,u}^- \leq |E| e^{-R} =: C_0 e^{-R}. \end{aligned} \tag{70}$$

Set

$$z_{u,v}^{(0)} := x_\ell^\top \mathbf{W}^{(0)} e_{u,v},$$

and, for $u \in \mathcal{U}_\ell$, $v \in \mathcal{N}(u)$,

$$\mathbf{W}_R e_{u,v} := \mathbf{W}^{(0)} e_{u,v} + \frac{z_{u,v}^{(R)} - z_{u,v}^{(0)}}{\|x_\ell\|_2^2} x_\ell,$$

with all other columns unchanged. Then

$$x_\ell^\top \mathbf{W}_R e_{u,v} = z_{u,v}^{(R)},$$
$$\|\mathbf{W}_R - \mathbf{W}^{(0)}\|_{\mathrm{F}}^2 = \frac{1}{\|x_\ell\|_2^2} \sum_{u \in \mathcal{U}_\ell} \sum_{v \in \mathcal{N}(u)} |z_{u,v}^{(R)} - z_{u,v}^{(0)}|^2.$$

Since

$$\|x_\ell\|_2^2 \geq 1,$$
$$r_{\ell,u}(v) \geq \frac{1}{|\mathcal{P}_\ell|} \quad \Longrightarrow \quad |z_{u,v}^{(R)}| \leq R + \log|\mathcal{P}_\ell|,$$
$$Z_{\ell,0} := \max_{\substack{u \in \mathcal{U}_\ell \\ v \in \mathcal{N}(u)}} |z_{u,v}^{(0)}| < \infty,$$

we obtain

$$\|\mathbf{W}_R - \mathbf{W}^{(0)}\|_{\mathrm{F}}^2 \leq |E|(R + \log|\mathcal{P}_\ell| + Z_{\ell,0})^2 \leq C_1(1 + R + \log|\mathcal{P}_\ell|)^2. \tag{71}$$

□

Write

$$\mathbf{g}_j = \frac{1}{B} \sum_{b=1}^{B} \mathbf{h}_j^{(b)}, \quad \boldsymbol{\zeta}_j := \mathbf{g}_j - \nabla F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(j-1)}),$$

where

$$\mathbf{h}_j^{(b)} := \mathbf{h}(\Gamma_j^{(b)}; \mathbf{W}_{\mathrm{SFT}}^{(j-1)}),$$
$$\boldsymbol{\xi}_j^{(b)} := \mathbf{h}_j^{(b)} - \mathbb{E}[\mathbf{h}_j^{(b)} \mid \mathcal{F}_{j-1}].$$

Then

$$\boldsymbol{\zeta}_j = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\xi}_j^{(b)}, \quad \mathbb{E}[\boldsymbol{\zeta}_j \mid \mathcal{F}_{j-1}] = 0.$$

Lemma 15 (Conditional mini-batch noise)**.** *The centered noise $\boldsymbol{\zeta}_j$ has conditional mean zero and variance at most G_ℓ^2/B. There is $\sigma_{\ell,B}^2 = O_G(B^{-1})$ such that Equation 73 holds for $0 \leq \lambda \leq \sigma_{\ell,B}^{-2}$.*

Proof. Conditional independence gives

$$\mathbb{E}[\|\boldsymbol{\zeta}_j\|_{\mathrm{F}}^2 \mid \mathcal{F}_{j-1}] = \frac{1}{B^2} \sum_{b=1}^{B} \mathbb{E}[\|\boldsymbol{\xi}_j^{(b)}\|_{\mathrm{F}}^2 \mid \mathcal{F}_{j-1}] \leq \frac{G_\ell^2}{B}. \tag{72}$$

Let $D := d|E|$. Since

$$\|\boldsymbol{\xi}_j^{(b)}\|_{\mathrm{F}} \leq 2G_\ell,$$

coordinate-wise Hoeffding gives, for every $t > 0$,

$$\mathbb{P}\left(\|\boldsymbol{\zeta}_j\|_{\mathrm{F}} \geq t \mid \mathcal{F}_{j-1}\right) \leq 2D \exp\left(-\frac{Bt^2}{8DG_\ell^2}\right).$$

Hence, for a constant $c_D > 0$ depending only on D,

$$\sigma_{\ell,B}^2 := \frac{c_D G_\ell^2}{B},$$
$$\mathbb{E}\left[\exp\left(\frac{\|\boldsymbol{\zeta}_j\|_{\mathrm{F}}^2}{\sigma_{\ell,B}^2}\right)\middle|\mathcal{F}_{j-1}\right] \le e.$$

For $0 \le \lambda \le \sigma_{\ell,B}^{-2}$, put $\theta := \lambda\sigma_{\ell,B}^2 \in [0,1]$. Then

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\|\boldsymbol{\zeta}_j\|_{\mathrm{F}}^2}\middle|\mathcal{F}_{j-1}\right] &= \mathbb{E}\left[\left(e^{\|\boldsymbol{\zeta}_j\|_{\mathrm{F}}^2/\sigma_{\ell,B}^2}\right)^\theta\middle|\mathcal{F}_{j-1}\right] \\ &\le \left(\mathbb{E}\left[e^{\|\boldsymbol{\zeta}_j\|_{\mathrm{F}}^2/\sigma_{\ell,B}^2}\middle|\mathcal{F}_{j-1}\right]\right)^\theta \\ &\le e^{\lambda\sigma_{\ell,B}^2}, \end{aligned} \tag{73}$$

and

$$\sigma_{\ell,B}^2 = O_G(B^{-1}).$$

□

Lemma 16 (Smoothness and a bounded initial transient)**.** *The potential is L_ℓ-smooth with $L_\ell \le d$. For $0 < \nu < 1$, there is a deterministic j_0, independent of n, B, δ, after which $\eta_j \le 1/(2L_\ell)$. The finite comparator of Lemma 14 satisfies Equation 75 uniformly over the training history before j_0.*

Proof. For a perturbation H, let

$$a_u(H) := x_\ell^\top H e_{u,\cdot}, \quad p_u := \pi_{\mathbf{W}}(\cdot \mid u, x_\ell).$$

From Equation (66),

$$\begin{aligned}\nabla^2 F_\ell(\mathbf{W})[H,H] &= \sum_{u\in\mathcal{U}_\ell} \omega_{\ell,u} a_u(H)^\top \left(\mathrm{Diag}(p_u) - p_u p_u^\top\right) a_u(H) \\ &\le \sum_{u\in\mathcal{U}_\ell} \omega_{\ell,u}\|a_u(H)\|_2^2 \le \|x_\ell\|_2^2\|H\|_{\mathrm{F}}^2 \le d\|H\|_{\mathrm{F}}^2.\end{aligned}$$

Thus

$$L_\ell \le d. \tag{74}$$

Since $\eta_j = \Theta(j^{-\nu})$, there exists a fixed j_0 such that

$$\eta_j \le \frac{1}{2L_\ell}, \qquad j \ge j_0.$$

For $n < j_0$, Equation (58) is made trivial by enlarging C_ℓ. Hence assume $n \ge j_0$. Also,

$$\|\mathbf{W}_{\mathrm{SFT}}^{(j_0-1)} - \mathbf{W}_R\|_{\mathrm{F}} \le \|\mathbf{W}^{(0)} - \mathbf{W}_R\|_{\mathrm{F}} + G_\ell \sum_{j=1}^{j_0-1} \eta_j,$$

so

$$\|\mathbf{W}_{\mathrm{SFT}}^{(j_0-1)} - \mathbf{W}_R\|_{\mathrm{F}}^2 \le C_2(1 + R + \log|\mathcal{P}_\ell|)^2. \tag{75}$$

□

Lemma 17 (A finite-comparator last-iterate bound)**.** *For every deterministic finite comparator $\mathbf{W}_R$ and $n \ge j_0$, the bound Equation 78 holds with probability at least $1-\delta$. Its constants are independent of n, B, δ, R.*

Proof. We next verify explicitly that the last-iterate argument does not require F_ℓ to attain its infimum at a finite parameter. Fix the finite comparator $\mathbf{W}_R$ constructed above. Equation 62 gives

$$2\eta_j \left\langle \mathbf{g}_j, \mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \mathbf{W}_R \right\rangle = \|\mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \mathbf{W}_R\|_{\mathrm{F}}^2 - \|\mathbf{W}_{\mathrm{SFT}}^{(j)} - \mathbf{W}_R\|_{\mathrm{F}}^2 + \eta_j^2 \|\mathbf{g}_j\|_{\mathrm{F}}^2. \tag{76}$$

Moreover, convexity of F_ℓ and the definition of $\boldsymbol{\zeta}_j$ imply

$$\begin{aligned} F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(j-1)}) - F_\ell(\mathbf{W}_R) &\le \left\langle \nabla F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(j-1)}), \mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \mathbf{W}_R \right\rangle \\ &= \left\langle \mathbf{g}_j, \mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \mathbf{W}_R \right\rangle - \left\langle \boldsymbol{\zeta}_j, \mathbf{W}_{\mathrm{SFT}}^{(j-1)} - \mathbf{W}_R \right\rangle. \end{aligned} \tag{77}$$

Equations 76 and 77 are the comparator-dependent one-step relations used in the Euclidean convex specialization of Lemma 4.3 of Liu & Zhou (2024). The subsequent weighted telescoping and martingale argument uses the reference point only through these relations and the finite initial Euclidean distance. Therefore, using Equation 73, the same argument applies with this arbitrary finite comparator $\mathbf{W}_R$ and does not require a finite minimizer of F_ℓ.

Specializing that argument to

$$\psi(\mathbf{W}) := \frac{1}{2}\|\mathbf{W}\|_{\mathrm{F}}^2, \quad \mu_f = \mu_h = M = 0, \quad \gamma_j = \eta_j,$$

thus gives, for every finite $\mathbf{W}_R$, with probability at least $1-\delta$,

$$F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(n)}) - F_\ell(\mathbf{W}_R) \le C_3 \left[\frac{\|\mathbf{W}_{\mathrm{SFT}}^{(j_0-1)} - \mathbf{W}_R\|_{\mathrm{F}}^2}{\sum_{j=j_0}^n \eta_j} + \sigma_{\ell,B}^2 \left(1 + 2\log\frac{2}{\delta}\right) \sum_{j=j_0}^n \frac{\eta_j^2}{\sum_{s=j}^n \eta_s} \right]. \tag{78}$$

□

Lemma 18 (Polynomial step-size sums)**.** *For fixed $0 < \nu < 1$ and $\eta_j = \Theta(j^{-\nu})$, $\sum_{j=j_0}^n \eta_j = \Theta(n^{1-\nu})$ for sufficiently large n and the weighted noise sum satisfies Equation 79.*

Proof.

$$\sum_{j=j_0}^n \eta_j = \Theta(n^{1-\nu}).$$

Moreover, with

$$S_n := \sum_{j=j_0}^n \frac{\eta_j^2}{\sum_{s=j}^n \eta_s} = S_n^{(1)} + S_n^{(2)},$$

where

$$S_n^{(1)} := \sum_{j=j_0}^{\lfloor n/2 \rfloor} \frac{\eta_j^2}{\sum_{s=j}^n \eta_s}, \qquad S_n^{(2)} := \sum_{j=\lfloor n/2 \rfloor + 1}^{n} \frac{\eta_j^2}{\sum_{s=j}^n \eta_s},$$

we have

$$\begin{aligned} S_n^{(1)} &\le \frac{C}{n^{1-\nu}} \sum_{j=1}^{\lfloor n/2 \rfloor} j^{-2\nu} \\ &= \begin{cases} O(n^{-\nu}), & 0 < \nu < 1/2, \\ O(\log(en) n^{-1/2}), & \nu = 1/2, \\ O(n^{-(1-\nu)}), & 1/2 < \nu < 1, \end{cases} \end{aligned}$$

and, for $k = n - j + 1$,

$$S_n^{(2)} \le C n^{-\nu} \sum_{k=1}^{\lceil n/2 \rceil} \frac{1}{k} = O(n^{-\nu} \log(en)).$$

Therefore

$$S_n = O\left(\frac{\log(en)}{n^{m_\nu}}\right). \tag{79}$$

□

Lemma 19 (Simultaneous SFT successor-set learning)**.** *For any finite initial parameter and fixed $n \geq 1$, the batch-explicit bound Equation 58 holds with probability at least $1-\delta$. No pretrained coverage condition is needed for this assertion.*

Proof. Apply Lemmas 14–18. Combining Equations (70), (73), (75), (78), and (79),

$$F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(n)}) \leq C_4 \left[e^{-R} + \frac{(1+R+\log|\mathcal{P}_\ell|)^2}{n^{1-\nu}} + \frac{(1+2\log(2/\delta))\log(en)}{B\, n^{m_\nu}} \right].$$

Choose

$$R := \log(e + n|\mathcal{P}_\ell|). \tag{80}$$

Then

$$e^{-R} = \frac{1}{e+n|\mathcal{P}_\ell|} \leq \frac{1}{n}, \qquad 1+R+\log|\mathcal{P}_\ell| \leq 3\log(e+n|\mathcal{P}_\ell|).$$

Hence

$$F_\ell(\mathbf{W}_{\mathrm{SFT}}^{(n)}) \leq C_5 \left[\frac{1}{n} + \frac{\log^2(e+n|\mathcal{P}_\ell|)}{n^{1-\nu}} + \frac{(1+2\log(2/\delta))\log(en)}{B\, n^{m_\nu}} \right]. \tag{81}$$

Lemma 13 converts this potential bound into Equation 58. The same scalar event controls every source, so no additional union bound is required. □

Main proof of Theorem 2.

Proof. By Lemma 19, Equation 58 holds with probability at least $1-\delta$. Since $0 < \nu < 1/2$, we have $m_\nu = \nu$. For $n \geq 1$ and $\delta \in (0,1)$,

$$\frac{1}{n} \leq \frac{1}{n^{1-\nu}}, \qquad 1 + 2\log\frac{2}{\delta} = O\Big(\log\frac{e}{\delta}\Big),$$

where the implied constant is independent of n, B, and δ. Absorbing the factors $\log^2(e+n|\mathcal{P}_\ell|)$ and $\log(en)$ into $\tilde{O}$, the three-term bound in Equation 58 gives Equation 13, retaining the stated dependence on B and δ. □

B.6 Retention under continual SD

Theorem 3 (No forgetting under SD)**.** *Let $\mathbf{W}_{\mathrm{SD}}^{[s]}$ denote the parameter after sequentially training $\mathcal{D}_1, \ldots, \mathcal{D}_s$ by SD. Fix $\ell \in \{1, \ldots, L\}$, and let n_ℓ be the number of SD updates used for $\mathcal{D}_\ell$. Suppose Assumptions 1 and 2 hold for all tasks, and the stepsize $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in [0, 1/2)$. Then, after any numbers of subsequent SD updates on $\mathcal{D}_{\ell+1}, \ldots, \mathcal{D}_L$, with probability at least $1-\delta$,*

$$\pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}\Big(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell\Big) \geq 1 - O(KM\epsilon_1) - O\Big(KM\xi_{\ell,n_\ell}^{(K)}(\delta)\Big). \tag{14}$$

Moreover, every correct transition learned for task $\mathcal{D}_\ell$ satisfies

$$\pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}(v \mid u, x_\ell) \geq \pi_{\mathbf{W}_{\mathrm{SD}}^{[\ell]}}(v \mid u, x_\ell), \qquad (u,v) \in E_\ell. \tag{15}$$

Sketch of proof of Theorem 3.

- Show that updates on other tasks preserve interior transitions and do not decrease correct transition probabilities, including before the target task is learned (Lemma 20).
- Apply the transition-learning argument conditionally on preceding-task training, using the preserved initialization bounds (Proposition 2 and Lemma 20).
- Preserve the learned transition bounds throughout all subsequent-task updates (Lemma 20).
- Convert the final transition bounds into old-task rollout accuracy on the same training event (Lemma 11).

Lemma 20 (Preservation of correct transitions across tasks)**.** *Under Assumption 1 and the nonnegative task inputs, an SD update on any other task preserves all interior-source logits and does not decrease any correct transition probability for task r. Therefore, after any sequence of SD updates, each correct task-r transition is at least its pretrained probability.*

Proof. Lemma 4 gives, for $u \in \mathring{C}_k$,

$$\pi_{\mathbf{W}^+}(\cdot \mid u, x_r) = \pi_{\mathbf{W}}(\cdot \mid u, x_r).$$

Fix a correct old transition $(u, v) \in E_r$. If a later update has no hint-source at u, all logits at u are unchanged. Otherwise, let (u, q) be a correct transition for the later task that contributes to the update at u, so that

$$\mathbf{W}^+ = \mathbf{W} + \gamma_{s,u} x_s \left(e_{u,q} - \sum_{z \in \mathcal{N}(u)} \pi_{\mathbf{W}}(z \mid u, x_s) e_{u,z} \right)^\top, \qquad \gamma_{s,u} \geq 0.$$

If $\langle x_r, x_s \rangle = 0$, then, for every $z \in \mathcal{N}(u)$,

$$x_r^\top (\mathbf{W}^+ - \mathbf{W}) e_{u,z} = 0.$$

Hence all task-r logits, and therefore the entire transition distribution at u, are unchanged, irrespective of whether $q = v$.

If $\langle x_r, x_s \rangle > 0$, Assumption 1 gives $q = v$. For $w \neq v$,

$$\begin{aligned} &x_r^\top \mathbf{W}^+ (e_{u,v} - e_{u,w}) - x_r^\top \mathbf{W} (e_{u,v} - e_{u,w}) \\ &= \gamma_{s,u} \langle x_r, x_s \rangle \left[1 - \pi_{\mathbf{W}}(v \mid u, x_s) + \pi_{\mathbf{W}}(w \mid u, x_s) \right] \geq 0, \end{aligned}$$

Hence

$$\pi_{\mathbf{W}^+}(v \mid u, x_r) \geq \pi_{\mathbf{W}}(v \mid u, x_r),$$

Applying these two cases to every later-task contribution and every subsequent SD update gives the exact monotonicity

$$\pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}(v \mid u, x_r) \geq \pi_{\mathbf{W}_{\mathrm{SD}}^{[r]}}(v \mid u, x_r), \qquad (u, v) \in E_r.$$

The proof uses no ordering property of r and s, so it also applies to updates before task r is learned. Iteration from $\mathbf{W}^{(0)}$ proves the final assertion. □

Main proof of Theorem 3.

Proof. Apply Lemma 20 to the tasks preceding ℓ. At the start of task ℓ, its interior kernel is still pretrained and its correct transition probabilities are at least $c_{\mathrm{int}} \epsilon_2$. The proof of Proposition 2, using Lemmas 8–10, therefore applies conditionally on the preceding history with the same uniform constants. After integrating out that history, an event of probability at least $1 - \delta$ gives $1 - p_i \leq C_{\mathrm{tr}} \xi_{\ell, n_\ell}^{(K)}(\delta)$ at task completion. Lemma 20 preserves all these inequalities during every later task. On that same event, Lemma 11 applied to the final predictor gives

$$1 - \pi_{\mathbf{W}_{\mathrm{SD}}^{[L]}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) \leq C_{\mathrm{bad}}(1 + N_\ell)\epsilon_1 + N_\ell C_{\mathrm{tr}} \xi_{\ell, n_\ell}^{(K)}(\delta).$$

Since $N_\ell \leq KM$, this is Equation 14. The edgewise assertion is exactly Lemma 20. □

B.7 Catastrophic forgetting under SFT

Notation for Theorem 4. For a path $\gamma = (u_0, \ldots, u_T)$, set

$$\mathrm{Src}(\gamma) := \{u_0, \ldots, u_{T-1}\}.$$

Define

$$\mathcal{U}_{\text{new}} := \bigcup_{\Gamma \in \mathcal{P}_{\text{new}}} \text{Src}(\Gamma), \tag{82}$$

$$\mathcal{A}_{\text{new}}(u) := \{v : \exists \Gamma \in \mathcal{P}_{\text{new}} \text{ such that } (u, v) \in \Gamma\}, \qquad u \in \mathcal{U}_{\text{new}}. \tag{83}$$

The conflicted old correct paths and their fraction are

$$\mathcal{C}_{\text{old}} := \{\gamma = (u_0, \ldots, u_T) \in \mathcal{P}_{\text{old}} : \exists t < T,\ u_t \in \mathcal{U}_{\text{new}},\ u_{t+1} \notin \mathcal{A}_{\text{new}}(u_t)\}, \tag{84}$$

$$\theta := \frac{|\mathcal{C}_{\text{old}}|}{|\mathcal{P}_{\text{old}}|}. \tag{85}$$

Let

$$P^{\text{old}}_{\text{pre}}(\gamma) := \pi_{\mathbf{W}^{(0)}_{\text{new}}}(Y_{0:T} = \gamma \mid Y_0 = y_{\text{old},0}, x_{\text{old}}), \tag{86}$$

$$\begin{aligned} p^{\text{pre}}_{\text{old}} &:= \pi_{\mathbf{W}^{(0)}_{\text{new}}}\left(Y_T \in \tilde{C}_{\text{old}} \mid Y_0 = y_{\text{old},0}, x_{\text{old}}\right) \\ &= \sum_{\gamma \in \mathcal{P}_{\text{old}}} P^{\text{old}}_{\text{pre}}(\gamma), \end{aligned} \tag{87}$$

$$R_{\text{pre}} := \sum_{\gamma \in \mathcal{P}_{\text{old}} \setminus \mathcal{C}_{\text{old}}} P^{\text{old}}_{\text{pre}}(\gamma), \tag{88}$$

$$M_{\text{old}} := |\mathcal{P}_{\text{old}}|.$$

The second equality in Equation 87 follows because the acyclic nature of the DAG makes every student rollout node-simple and $\mathcal{P}_{\text{old}}$ contains every node-simple path ending in $\tilde{C}_{\text{old}}$. For $s \in [0, 1]$, define

$$\mathfrak{C}_{\text{old}}(s) := \max_{\substack{\mathcal{S} \subseteq \mathcal{P}_{\text{old}} \\ |\mathcal{S}| \le \lfloor s M_{\text{old}} \rfloor}} \sum_{\gamma \in \mathcal{S}} P^{\text{old}}_{\text{pre}}(\gamma). \tag{89}$$

Set

$$\lambda := \frac{\langle x_{\text{old}}, x_{\text{new}} \rangle}{\|x_{\text{new}}\|_2^2} > 0, \quad m_\nu := \min\{\nu, 1 - \nu\}. \tag{90}$$

For $u \in \mathcal{U}_{\text{new}}$ and $v \in \mathcal{N}(u)$, define

$$b_{u,v} := x_{\text{old}}^\top \mathbf{W}^{(0)}_{\text{new}} e_{u,v} - \lambda x_{\text{new}}^\top \mathbf{W}^{(0)}_{\text{new}} e_{u,v}, \tag{91}$$

$$b_{\min}(u) := \min_{w \in \mathcal{N}(u)} b_{u,w},$$

$$\Delta_{\text{map}} := \max_{u \in \mathcal{U}_{\text{new}}} \left[\max_{v \in \mathcal{N}(u)} b_{u,v} - b_{\min}(u) \right], \tag{92}$$

$$C_{\text{map}} := e^{\Delta_{\text{map}}} = \max_{u \in \mathcal{U}_{\text{new}}} \max_{v \in \mathcal{N}(u)} e^{b_{u,v} - b_{\min}(u)}. \tag{93}$$

Let C_{new} be the constant from the batch-explicit bound of Lemma 19 applied to $\mathcal{D}_{\text{new}}$ from $\mathbf{W}^{(0)}_{\text{new}}$, and define

$$\bar{\epsilon}_n(\delta) := C_{\text{new}} |\mathcal{P}_{\text{new}}| \left[\frac{1}{n} + \frac{\log^2(e + n|\mathcal{P}_{\text{new}}|)}{n^{1-\nu}} + \frac{(1 + 2\log(2/\delta))\log(en)}{B n^{m_\nu}} \right]. \tag{94}$$

Finally, define the relative largest and smallest old-path masses by

$$\kappa^{\text{old}}_{\max} := \frac{M_{\text{old}}}{p^{\text{pre}}_{\text{old}}} \max_{\gamma \in \mathcal{P}_{\text{old}}} P^{\text{old}}_{\text{pre}}(\gamma), \quad \kappa^{\text{old}}_{\min} := \frac{M_{\text{old}}}{p^{\text{pre}}_{\text{old}}} \min_{\gamma \in \mathcal{P}_{\text{old}}} P^{\text{old}}_{\text{pre}}(\gamma). \tag{95}$$

Write

$$\mathbf{W}^{(j)} := \mathbf{W}^{(j)}_{\text{new}}.$$

For an old path $\gamma = (u_0, \ldots, u_T)$, set

$$P^{\text{old}}_j(\gamma) := \pi_{\mathbf{W}^{(j)}}(Y_{0:T} = \gamma \mid Y_0 = y_{\text{old},0}, x_{\text{old}}).$$

Theorem 4 (Catastrophic forgetting under SFT)**.** *Fix $\delta \in (0,1)$ and set $\lambda := \frac{\langle x_{\mathrm{old}}, x_{\mathrm{new}} \rangle}{\|x_{\mathrm{new}}\|_2^2} > 0$. Suppose $\eta_j = \Theta(j^{-\nu})$ for some $\nu \in (0, 1/2)$. Then, for arbitrary pre-new-task old-path masses, with probability at least $1-\delta$,*

$$\Pi(\mathbf{W}_{\mathrm{new}}^{(n)}) \leq \underbrace{\mathfrak{C}_{\mathrm{old}}(1-\theta)}_{\substack{\text{upper bound on} \\ \text{surviving knowledge}}} + \widetilde{O}\left(\theta |\mathcal{P}_{\mathrm{old}}||\mathcal{P}_{\mathrm{new}}|^{\lambda} \left[\frac{1 + B^{-1}\log(1/\delta)}{n^{\nu}}\right]^{\lambda}\right), \tag{16}$$

$$\Pi(\mathbf{W}_{\mathrm{new}}^{(0)}) - \mathfrak{C}_{\mathrm{old}}(\theta) \leq \Pi(\mathbf{W}_{\mathrm{new}}^{(n)}), \tag{17}$$

where $\Pi(\mathbf{W}) = \pi_{\mathbf{W}}(Y_T \in \tilde{C}_{\mathrm{old}} \mid Y_0 = y_{\mathrm{old},0}, x_{\mathrm{old}})$ and

$$\mathfrak{C}_{\mathrm{old}}(\phi) := \max_{\substack{\mathcal{P} \subseteq \mathcal{P}_{\mathrm{old}} \\ |\mathcal{P}| \leq \phi |\mathcal{P}_{\mathrm{old}}|}} \sum_{\gamma \in \mathcal{P}} \pi_{\mathbf{W}_{\mathrm{new}}^{(0)}}(Y_{0:T} = \gamma \mid Y_0 = y_{\mathrm{old},0}, x_{\mathrm{old}}). \tag{18}$$

$\tilde{O}(\cdot)$ hides logarithmic factors in n and $|\mathcal{P}_{\mathrm{new}}|$.

Sketch of proof of Theorem 4.

- Characterize conflicted old paths by their overlap with new-task source nodes (Lemma 21).
- Show that new-task SFT preserves every non-conflicted old-path mass exactly (Lemma 22).
- Bound the probability of unsupported successors under the new-task input (Lemma 19).
- Transfer successor suppression from the new-task input to the old-task input using the exact logit relation (Lemma 23).
- Sum the preserved and suppressed path masses to obtain the old-task retention bounds (Lemmas 22 and 24).

Lemma 21 (Support separation equals source overlap)**.** *Under the disjoint terminal-cluster assumptions, an old correct path belongs to $\mathcal{C}_{\mathrm{old}}$ if and only if its source set intersects $\mathcal{U}_{\mathrm{new}}$. Thus Equation 96 and Equation 97 hold.*

Proof. Let

$$\mathcal{S}_{\mathrm{old}} := \{\gamma \in \mathcal{P}_{\mathrm{old}} : \mathrm{Src}(\gamma) \cap \mathcal{U}_{\mathrm{new}} \neq \varnothing\}.$$

By Equation 84,

$$\mathcal{C}_{\mathrm{old}} \subseteq \mathcal{S}_{\mathrm{old}}.$$

Conversely, fix

$$\gamma = (u_0, \ldots, u_T) \in \mathcal{S}_{\mathrm{old}}$$

and let

$$t_\star := \max\{t < T : u_t \in \mathcal{U}_{\mathrm{new}}\}.$$

Suppose

$$u_{t_\star+1} \in \mathcal{A}_{\mathrm{new}}(u_{t_\star}).$$

Then there exists $\Gamma \in \mathcal{P}_{\mathrm{new}}$ containing $(u_{t_\star}, u_{t_\star+1})$. If $u_{t_\star+1}$ is not the final node of Γ, then

$$u_{t_\star+1} \in \mathcal{U}_{\mathrm{new}},$$

contradicting the maximality of $t_\star$. If $u_{t_\star+1}$ is the final node of Γ, then

$$u_{t_\star+1} \in \widetilde{C}_{\mathrm{new}}.$$

Since $\widetilde{C}_{\mathrm{new}}$ has no outgoing inter-cluster transition and

$$\widetilde{C}_{\mathrm{old}} \cap \widetilde{C}_{\mathrm{new}} = \varnothing,$$

the old correct path γ cannot subsequently reach $\widetilde{C}_{\mathrm{old}}$, again a contradiction. Therefore

$$u_{t_\star+1} \notin \mathcal{A}_{\mathrm{new}}(u_{t_\star}),$$

so $\gamma \in \mathcal{C}_{\mathrm{old}}$. Hence

$$\mathcal{C}_{\mathrm{old}} = \{\gamma \in \mathcal{P}_{\mathrm{old}} : \mathrm{Src}(\gamma) \cap \mathcal{U}_{\mathrm{new}} \neq \varnothing\}. \tag{96}$$

Consequently,

$$\gamma \notin \mathcal{C}_{\mathrm{old}} \quad \Longrightarrow \quad \mathrm{Src}(\gamma) \cap \mathcal{U}_{\mathrm{new}} = \varnothing. \tag{97}$$

□

Lemma 22 (Exact preservation of non-conflicted path mass)**.** *For every training history and $\gamma \notin \mathcal{C}_{\mathrm{old}}$, $P_n^{\mathrm{old}}(\gamma) = P_{\mathrm{pre}}^{\mathrm{old}}(\gamma)$. Moreover, $p_{\mathrm{old}}^{\mathrm{pre}} - \mathfrak{C}_{\mathrm{old}}(\theta) \le R_{\mathrm{pre}} \le \mathfrak{C}_{\mathrm{old}}(1-\theta)$.*

Proof. Lemma 12 changes only edge columns whose source node lies in $\mathcal{U}_{\mathrm{new}}$. Thus Equation 97 gives, for every $\gamma \notin \mathcal{C}_{\mathrm{old}}$,

$$P_n^{\mathrm{old}}(\gamma) = P_0^{\mathrm{old}}(\gamma) = P_{\mathrm{pre}}^{\mathrm{old}}(\gamma). \tag{98}$$

Consequently,

$$\sum_{\gamma \notin \mathcal{C}_{\mathrm{old}}} P_n^{\mathrm{old}}(\gamma) = R_{\mathrm{pre}}. \tag{99}$$

Since $|\mathcal{P}_{\mathrm{old}} \setminus \mathcal{C}_{\mathrm{old}}| = (1-\theta)M_{\mathrm{old}}$, the concentration profile gives

$$R_{\mathrm{pre}} \le \mathfrak{C}_{\mathrm{old}}(1-\theta). \tag{100}$$

Also, $|\mathcal{C}_{\mathrm{old}}| = \theta M_{\mathrm{old}}$ implies

$$\sum_{\gamma \in \mathcal{C}_{\mathrm{old}}} P_{\mathrm{pre}}^{\mathrm{old}}(\gamma) \le \mathfrak{C}_{\mathrm{old}}(\theta), \tag{101}$$

and therefore

$$R_{\mathrm{pre}} = p_{\mathrm{old}}^{\mathrm{pre}} - \sum_{\gamma \in \mathcal{C}_{\mathrm{old}}} P_{\mathrm{pre}}^{\mathrm{old}}(\gamma) \ge p_{\mathrm{old}}^{\mathrm{pre}} - \mathfrak{C}_{\mathrm{old}}(\theta). \tag{102}$$

Since all path probabilities are nonnegative, Equations 99 and 102 prove the late-phase lower bound. □

Lemma 23 (Exact transfer between task inputs)**.** *For every training history, source $u \in \mathcal{U}_{\mathrm{new}}$, successor v, and iteration j, the probability identity Equation 104 holds. In particular,*

$$\pi_{\mathbf{W}^{(j)}}(v \mid u, x_{\mathrm{old}}) \le C_{\mathrm{map}} \pi_{\mathbf{W}^{(j)}}(v \mid u, x_{\mathrm{new}})^{\lambda}.$$

Proof. For $u \in \mathcal{U}_{\mathrm{new}}$ and $v \in \mathcal{N}(u)$, define

$$z_{i,u,v}^{(j)} := x_i^{\top} \mathbf{W}^{(j)} e_{u,v}, \qquad i \in \{\mathrm{old}, \mathrm{new}\}.$$

Every update of the column $e_{u,v}$ during new-task SFT lies in $\mathrm{span}\{x_{\mathrm{new}}\}$. Hence

$$z_{\mathrm{old},u,v}^{(j)} - z_{\mathrm{old},u,v}^{(j-1)} = \lambda \left(z_{\mathrm{new},u,v}^{(j)} - z_{\mathrm{new},u,v}^{(j-1)} \right),$$

and Equation 91 gives

$$z_{\mathrm{old},u,v}^{(j)} = b_{u,v} + \lambda z_{\mathrm{new},u,v}^{(j)}. \tag{103}$$

Writing

$$p_{u,v}^{(j)} := \pi_{\mathbf{W}^{(j)}}(v \mid u, x_{\mathrm{new}}),$$

Equation 103 implies

$$\pi_{\mathbf{W}^{(j)}}(v \mid u, x_{\text{old}}) = \frac{e^{b_{u,v}}(p_{u,v}^{(j)})^{\lambda}}{\sum_{w \in \mathcal{N}(u)} e^{b_{u,w}}(p_{u,w}^{(j)})^{\lambda}}. \tag{104}$$

Since $x_{\text{old}}, x_{\text{new}} \in \{0,1\}^d$,

$$0 < \lambda \leq 1, \qquad \sum_{w \in \mathcal{N}(u)} (p_{u,w}^{(j)})^{\lambda} \geq \sum_{w \in \mathcal{N}(u)} p_{u,w}^{(j)} = 1.$$

Therefore the denominator in Equation 104 satisfies

$$\sum_{w \in \mathcal{N}(u)} e^{b_{u,w}}(p_{u,w}^{(j)})^{\lambda} \geq e^{b_{\min}(u)} \sum_{w \in \mathcal{N}(u)} (p_{u,w}^{(j)})^{\lambda} \geq e^{b_{\min}(u)}. \tag{105}$$

For $i \in \{\text{old}, \text{new}\}$, put

$$p_{i,u,v}^{(0)} := \pi_{\mathbf{W}_{\text{new}}^{(0)}}(v \mid u, x_i).$$

For every $u \in \mathcal{U}_{\text{new}}$ and $v, w \in \mathcal{N}(u)$,

$$e^{b_{u,v} - b_{u,w}} = \frac{p_{\text{old},u,v}^{(0)}}{p_{\text{old},u,w}^{(0)}} \left(\frac{p_{\text{new},u,w}^{(0)}}{p_{\text{new},u,v}^{(0)}} \right)^{\lambda}. \tag{106}$$

Consequently,

$$C_{\text{map}} = e^{\Delta_{\text{map}}} = \max_{u \in \mathcal{U}_{\text{new}}} \max_{v,w \in \mathcal{N}(u)} \frac{p_{\text{old},u,v}^{(0)}}{p_{\text{old},u,w}^{(0)}} \left(\frac{p_{\text{new},u,w}^{(0)}}{p_{\text{new},u,v}^{(0)}} \right)^{\lambda}. \tag{107}$$

Combining Equations 104 and 105,

$$\pi_{\mathbf{W}^{(j)}}(v \mid u, x_{\text{old}}) \leq e^{b_{u,v} - b_{\min}(u)} (p_{u,v}^{(j)})^{\lambda} \leq C_{\text{map}} (p_{u,v}^{(j)})^{\lambda}. \tag{108}$$

□

Lemma 24 (From successor suppression to total old-path mass)**.** *Fix any training history satisfying Equation 113. Then*

$$R_{\text{pre}} \leq \Pi(\mathbf{W}^{(n)}) \leq R_{\text{pre}} + C_{\text{map}} \theta M_{\text{old}} \bar{\epsilon}_n (\delta')^{\lambda}.$$

Here the premise is a deterministic property of the fixed history; it is supplied with high probability by Lemma 19.

Proof. For every $\gamma = (u_0, \ldots, u_T) \in \mathcal{C}_{\text{old}}$, Equation 84 gives an index $t_\gamma < T$ such that

$$u_{t_\gamma} \in \mathcal{U}_{\text{new}}, \quad u_{t_\gamma + 1} \notin \mathcal{A}_{\text{new}}(u_{t_\gamma}).$$

Thus Lemma 23 yields

$$P_n^{\text{old}}(\gamma) \leq C_{\text{map}} \bar{\epsilon}_n (\delta')^{\lambda}. \tag{109}$$

Since $|\mathcal{C}_{\text{old}}| = \theta M_{\text{old}}$,

$$\sum_{\gamma \in \mathcal{C}_{\text{old}}} P_n^{\text{old}}(\gamma) \leq C_{\text{map}} \theta M_{\text{old}} \bar{\epsilon}_n (\delta')^{\lambda}. \tag{110}$$

Because the student rollout is node-simple and $\mathcal{P}_{\text{old}}$ contains all node-simple correct paths, its terminal-cluster probability is exactly

$$\pi_{\mathbf{W}^{(n)}} \Big(Y_T \in \tilde{C}_{\text{old}} \mid Y_0 = y_{\text{old},0}, x_{\text{old}} \Big) = \sum_{\gamma \in \mathcal{P}_{\text{old}}} P_n^{\text{old}}(\gamma). \tag{111}$$

Together with Equation 99, this gives

$$\begin{aligned} R_{\text{pre}} &\leq \pi_{\mathbf{W}^{(n)}} \Big(Y_T \in \tilde{C}_{\text{old}} \mid Y_0 = y_{\text{old},0}, x_{\text{old}} \Big) \\ &\leq R_{\text{pre}} + C_{\text{map}} \theta M_{\text{old}} \bar{\epsilon}_n (\delta')^{\lambda}. \end{aligned} \tag{112}$$

□

Main proof of Theorem 4.

Proof. Fix $\delta' \in (0,1)$. Apply Lemma 19 to $\mathcal{D}_{\text{new}}$ with initial point $\mathbf{W}_{\text{new}}^{(0)}$. With probability at least $1-\delta'$,

$$\max_{u \in \mathcal{U}_{\text{new}}} \left[1 - \pi_{\mathbf{W}^{(n)}}(\mathcal{A}_{\text{new}}(u) \mid u, x_{\text{new}})\right] \leq \bar{\epsilon}_n(\delta'), \tag{113}$$

where $\bar{\epsilon}_n(\delta')$ is defined in Equation 94 and satisfies

$$\bar{\epsilon}_n(\delta') = \widetilde{O}_{G,\mathbf{W}_{\text{new}}^{(0)}}\left(|\mathcal{P}_{\text{new}}|\left[\frac{1}{n^{1-\nu}} + \frac{\log(1/\delta')}{Bn^{m_\nu}}\right]\right) \longrightarrow 0.$$

By Lemmas 23 and 24, this event also gives Equation 112. Since $\nu \in (0, 1/2)$, Equation 94 implies

$$C_{\text{map}}\theta M_{\text{old}}\bar{\epsilon}_n(\delta')^\lambda = \widetilde{O}\left(C_{\text{map}}\theta M_{\text{old}} C_{\text{new}}^\lambda |\mathcal{P}_{\text{new}}|^\lambda \left[\frac{1 + B^{-1}\log(1/\delta')}{n^\nu}\right]^\lambda\right). \tag{114}$$

For the late-phase statement, take $\delta' = \delta$ and combine Equations 112, 100, and 102.

If $\theta = 0$, Equations 96 and 98 give

$$\pi_{\mathbf{W}^{(n)}}\left(Y_T \in \tilde{C}_{\text{old}} \mid Y_0 = y_{\text{old},0}, x_{\text{old}}\right) = p_{\text{old}}^{\text{pre}}. \tag{115}$$

□

Remark 1 (Explicit representation of $\mathfrak{C}_{\text{old}}(\phi)$)**.** *Let $\kappa_{\max}^{\text{old}}$ and $\kappa_{\min}^{\text{old}}$ be the largest and smallest old-path masses normalized by their average mass, as defined in Appendix B.7. Then*

$$\mathfrak{C}_{\text{old}}(\phi) \leq \Pi(\mathbf{W}_{\text{new}}^{(0)}) \min\left\{\kappa_{\max}^{\text{old}}\phi, 1 - \kappa_{\min}^{\text{old}} + \kappa_{\min}^{\text{old}}\phi\right\}. \tag{19}$$

Sketch of proof of Remark 1.

- Bound the mass of the largest admissible subset using the largest individual old-path mass.
- Obtain a second bound by subtracting the smallest omitted masses from the total, and combine the two bounds.

Main proof of Remark 1.

Proof. Let

$$p_{(1)}^{\text{old}} \geq \cdots \geq p_{(M_{\text{old}})}^{\text{old}}$$

be the ordered values of $\{P_{\text{pre}}^{\text{old}}(\gamma) : \gamma \in \mathcal{P}_{\text{old}}\}$. For $\phi \in [0,1]$, put $q = \lfloor \phi M_{\text{old}} \rfloor$. Then

$$\mathfrak{C}_{\text{old}}(\phi) = \sum_{i=1}^{q} p_{(i)}^{\text{old}} \leq q \max_{\gamma \in \mathcal{P}_{\text{old}}} P_{\text{pre}}^{\text{old}}(\gamma) \leq p_{\text{old}}^{\text{pre}} \kappa_{\max}^{\text{old}} \phi. \tag{116}$$

Moreover, $M_{\text{old}} - q \geq (1-\phi)M_{\text{old}}$, so

$$\mathfrak{C}_{\text{old}}(\phi) = p_{\text{old}}^{\text{pre}} - \sum_{i=q+1}^{M_{\text{old}}} p_{(i)}^{\text{old}} \leq p_{\text{old}}^{\text{pre}}\left[1 - \kappa_{\min}^{\text{old}}(1-\phi)\right]. \tag{117}$$

Combining Equations 116 and 117 gives

$$\mathfrak{C}_{\text{old}}(\phi) \leq p_{\text{old}}^{\text{pre}} \min\left\{\kappa_{\max}^{\text{old}}\phi, 1 - \kappa_{\min}^{\text{old}} + \kappa_{\min}^{\text{old}}\phi\right\}. \tag{118}$$

□

B.8 PRE-TRAINING AND INTERMEDIATE RESTARTABILITY

Proposition 3 (Uniform restartability)**.** *Suppose Assumptions 1 and 2 hold, $\eta_j = \Theta(j^{-\nu})$ for fixed $0 \le \nu < 1/2$, and $0 < \epsilon_1, \epsilon_2, \delta, \epsilon_{\mathrm{rel}} < 1$. If n satisfies Equation 20 with a sufficiently large uniform constant, then with probability at least $1-\delta$ over training,*

$$\inf_{h:\mathcal{S}_{\ell,h}\neq\varnothing} \inf_{s\in\mathcal{S}_{\ell,h}} \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell) \ge 1 - O(KM\epsilon_1), \tag{119}$$

where $\mathcal{S}_{\ell,h}$ is the set of relevant starts in C_h.

Proposition 1 (Restart failure beyond a rare entrance)**.** *Suppose Assumptions 1 and 3 hold, $\eta_j = \Theta(j^{-\nu})$ for fixed $\nu \in [0, 1/2)$, and $0 < \epsilon_1, \epsilon_2, \epsilon_{\mathrm{rel}}, \delta < 1$. If*

$$n = \Omega\left(\max\left\{\frac{\log(2KM/\delta)}{B\underline{\rho}^2_{\ell,K,M}}\,,\ \left[\frac{\epsilon_1^{-1}-1}{\epsilon_2^2\log(1/\epsilon_{\mathrm{rel}})\underline{\rho}_{\ell,K,M}}\right]^{1/(1-\nu)}\right\}\right), \tag{20}$$

$$M' = \Omega\left(\frac{1}{\epsilon_2^{|E_\ell|}}\max\left\{\frac{1}{KM\epsilon_1}, \frac{Bn}{\delta}\right\}\right), \tag{21}$$

with sufficiently large uniform constants, then, with probability at least $1-\delta$, no training rollout visits u^*, and jointly

$$\pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) = 1 - O(KM\epsilon_1) \quad \text{(Misleading high accuracy)}, \tag{22}$$

$$\sup_{s\in\mathcal{S}_{\ell,k}} \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell) = O(M\epsilon_2) \quad \text{(Restart failure)}. \tag{23}$$

Here $\mathcal{S}_{\ell,k}$ contains the relevant starts in C_k and $\underline{\rho}_{\ell,K,M}$ is specified in Appendix B.3.

Sketch of proofs of Propositions 3 and 1.

- Show that the stated training length implies the effective-drift bound required for transition learning (Lemma 25).
- Under full coverage, learn all correct transitions on a common event and convert their accuracy into successful reasoning from every relevant start (Lemmas 10 and 11).
- Under the rare-entrance condition, bound the increase in visitation probability throughout training (Lemma 26).
- Use this visitation bound to establish nonvisitation across all adaptively generated training rollouts (Lemma 27).
- Learn the transitions reachable without the rare entrance and bound the final error from the original start (Lemmas 9, 10, and 11).
- Show that nonvisitation leaves the downstream cluster unchanged and bounds its restart accuracy (Lemma 28).
- Intersect the learning and nonvisitation events to obtain the joint original-start and restart guarantees (Proposition 1).

Lemma 25 (The stated training length implies the drift condition)**.** *Under the learning-rate conditions of Proposition 3, Equation 20 with a sufficiently large uniform constant implies*

$$\mathfrak{D}^{(K)}_{\ell,n}(\delta/2) \ge \frac{\epsilon_1^{-1}-1}{c^2_{\mathrm{int}}\epsilon_2^2\log(1/\epsilon_{\mathrm{rel}})}. \tag{120}$$

Proof. For fixed $0 \le \nu < 1/2$ and uniform learning-rate constants,

$$\sum_{j=1}^{n}\eta_j = \Theta(n^{1-\nu}), \qquad \sum_{j=1}^{n}\eta_j^2 = \Theta(n^{1-2\nu}).$$

The ratio of the square-root term in Equation 47 to its positive drift term is

$$O\left(\sqrt{\frac{\log(2KM/\delta)}{B\underline{\rho}^2_{\ell,K,M}n}}\right).$$

The first term of the maximum in Equation 20, with a sufficiently large constant, makes this ratio at most one half. Consequently the effective drift is $\Omega(\underline{\rho}_{\ell,K,M} n^{1-\nu})$. The second term in that maximum, again with a sufficiently large constant, makes this lower bound at least the right-hand side of Equation 120. This proves the implication. □

Lemma 26 (Control of the increase in visitation)**.** *Let $N = N_\ell = |E_\ell|$, the total number of correct inter-cluster edges of the task, be fixed. Under the retained initialization condition, for every training history and iteration j,*

$$\begin{aligned}&\pi_{\mathbf{W}^{(j)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\\ &\quad\le \pi_{\mathbf{W}^{(0)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\frac{c_{\mathrm{int}}^{-N}}{\epsilon_2^N}.\end{aligned} \tag{121}$$

Proof. Iterating the second inequality of Equation 51 from 1 to j gives

$$\frac{p_{u_i,v}^{(j)}}{p_{u_i,v}^{(0)}} \le \frac{p_i^{(j)}}{p_i^{(0)}} \le \frac{1}{p_i^{(0)}} \qquad ((u_i, v) \in E).$$

This applies to every outgoing edge at an updated source, including intra-cluster edges. At other sources the ratio is one. Factor the probability of a complete rollout into its one-step Markov probabilities, using the same stopping rule for both predictors. Each updated source occurs at most once by node simplicity. Sources that do not occur can be included because $p_i^{(j)}/p_i^{(0)} \ge 1$. Consequently,

$$\pi_{\mathbf{W}^{(j)}}(y_{0:T} \mid y_{\ell,0}, x_\ell) \le \pi_{\mathbf{W}^{(0)}}(y_{0:T} \mid y_{\ell,0}, x_\ell) \prod_{i=1}^{N_\ell} \frac{p_i^{(j)}}{p_i^{(0)}}. \tag{122}$$

The retained initialization condition gives $p_i^{(0)} \ge c_{\mathrm{int}}\epsilon_2$, and hence

$$\prod_{i=1}^{N_\ell} \frac{1}{p_i^{(0)}} \le c_{\mathrm{int}}^{-N}\epsilon_2^{-N}. \tag{123}$$

Sum Equation 122 over the node-simple rollouts that visit u_m before stopping, and use $p_i^{(j)} \le 1$ together with Equation 123. □

Lemma 27 (Nonvisitation throughout adaptive training)**.** *Under Assumption 3,*

$$\begin{aligned}&\mathbb{P}\left(\sum_{j=1}^{n}\sum_{b=1}^{B} \mathbf{1}\{\tau_m^{(b,j)} < T^{(b,j)}\} > 0\right)\\ &\quad\le Bn\,\pi_{\mathbf{W}^{(0)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\frac{c_{\mathrm{int}}^{-N}}{\epsilon_2^N} = O\left(\frac{Bn}{M'\epsilon_2^N}\right).\end{aligned} \tag{124}$$

Proof. Conditional on $\mathcal{F}_{j-1}$, the student rollout has law $\pi_{\mathbf{W}^{(j-1)}}$. Lemma 26 therefore gives

$$\begin{aligned}&\mathbb{E}[\mathbf{1}\{\tau_m^{(b,j)} < T^{(b,j)}\} \mid \mathcal{F}_{j-1}]\\ &\quad= \pi_{\mathbf{W}^{(j-1)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\\ &\quad\le \pi_{\mathbf{W}^{(0)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\frac{c_{\mathrm{int}}^{-N}}{\epsilon_2^N}.\end{aligned}$$

Here $T^{(b,j)}$ is the random length of that sampled rollout. Take total expectations and sum over b, j. The probability of a positive visit count is at most its expectation, proving the claim. □

Lemma 28 (An unvisited downstream cluster remains unlearned)**.** *Under Assumptions 1 and 3, if no training rollout visits u_m, then $\mathbf{W}^{(n)}e_{u,v} = \mathbf{W}^{(0)}e_{u,v}$ for every $u \in C_k$ and $(u, v) \in E$, and Equation 23 holds.*

Proof. Every training rollout starts outside C_k. Its only entrance is e_m, so nonvisitation of u_m implies nonvisitation of all of C_k. Lemma 7, iterated over the training steps, therefore leaves all columns with source in C_k unchanged, even if a hint contains an edge inside C_k.

Success from $s \in \mathcal{S}_{\ell,k}$ requires leaving $C_k \neq \tilde{C}_\ell$. Since s is a relevant start, a correct path from $y_{\ell,0}$ has a prefix ending at s. Concatenating this prefix with any successful continuation from s gives a correct path: a repeated vertex would create a directed cycle. Consequently, the first exit of a successful continuation belongs to E_ℓ. By Assumption 1, at most one such exit leaves each $u \in \partial^\star_\ell C_k$, so there are at most $|\partial^\star_\ell C_k| \leq M$ such edges. For each $(u, v) \in E_\ell$ with $u \in C_k$, the Markov property gives

$$\mathbb{E}_{\mathbf{W}^{(n)}}\left[\sum_{t<T} \mathbf{1}\{y_t = u, y_{t+1} = v\} \mid y_0 = s, x_\ell\right]$$
$$= \pi_{\mathbf{W}^{(n)}}(v \mid u, x_\ell)\mathbb{E}_{\mathbf{W}^{(n)}}\left[\sum_{t<T} \mathbf{1}\{y_t = u\} \mid y_0 = s, x_\ell\right] \leq \pi_{\mathbf{W}^{(0)}}(v \mid u, x_\ell).$$

The last step uses parameter preservation and node simplicity. A union bound over these correct exit edges and the retained initialization condition yield

$$\pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell)$$
$$\leq \sum_{\substack{(u,v)\in E_\ell \\ u\in C_k}} \pi_{\mathbf{W}^{(0)}}(v \mid u, x_\ell) = O(|\partial^\star_\ell C_k|\epsilon_2) = O(M\epsilon_2).$$

The bound is uniform over $s \in \mathcal{S}_{\ell,k}$, proving Equation 23. □

Main proofs of Propositions 3 and 1.

Proof of Proposition 3. Lemma 25 first gives Equation 120 from the stated condition on n. By Lemma 10 with failure budget $\delta/2$ and Equation 120, all N_ℓ correct edges have error at most ϵ_1 with probability at least $1 - \delta/2$. Fix any such training history. Lemma 11 gives, for every relevant start,

$$1 - \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_0 = s, x_\ell) \leq [C_{\mathrm{bad}}(1 + N_\ell) + N_\ell]\epsilon_1 \leq (1 + 2C_{\mathrm{bad}})KM\epsilon_1.$$

The same history works for every s, so no additional probabilistic union bound is needed. This is stronger than the stated confidence $1 - \delta$. □

Proof of Proposition 1. The common condition on n implies Equation 120 by Lemma 25. Hold $N = N_\ell = |E_\ell|$ fixed. Let C/M' be the initial visitation bound in Assumption 3, and take the uniform constant in Equation 21 to be at least $2Cc_{\mathrm{int}}^{-N}$. Then Equation 21 and initial rarity imply

$$\pi_{\mathbf{W}^{(0)}}(\exists t < T : y_t = u_m \mid y_{\ell,0}, x_\ell)\frac{c_{\mathrm{int}}^{-N}}{\epsilon_2^N} \leq \min\left\{KM\epsilon_1, \frac{\delta}{2Bn}\right\}. \tag{125}$$

Lemma 27 makes the probability that any training rollout visits u_m at most $\delta/2$. Independently of whether nonvisitation holds, Lemma 10 with failure budget $\delta/2$ and Equation 120 give error at most ϵ_1 for all edges specified in Lemma 9, except on a set of probability at most $\delta/2$. A union bound gives both conclusions jointly with probability at least $1-\delta$; the two events need not be independent.

Fix a history in their intersection. Lemma 26 and Equation 125 bound the final evaluation visitation probability by $KM\epsilon_1$. In Equation 57 this yields

$$1 - \pi_{\mathbf{W}^{(n)}}(y_T \in \tilde{C}_\ell \mid y_{\ell,0}, x_\ell) \leq [C_{\mathrm{bad}}(1 + KM) + 2KM]\epsilon_1 \leq (2 + 2C_{\mathrm{bad}})KM\epsilon_1.$$

Lemma 28 gives the restart upper bound on that same history. This proves all three assertions, including the nonvisit event, within confidence $1 - \delta$. □

Remark 3 (Separation of training times)**.** *Under Assumptions 1 and 2, the model trained by SD can resume from any relevant start at the same training length Equation 20 (Proposition 3 in Appendix B.8.) Holding other quantities fixed, the nonvisit horizon grows linearly with M'. If the nonvisit horizon grows asymptotically faster than the sufficient learning scale, there is a range of training lengths for which the two settings exhibit different restart behavior. Under the full-coverage condition, the trained model can reason successfully from every relevant start, whereas under the rare-entrance condition, the cluster beyond the rare entrance remains unlearned.*

Sketch of proof of Remark 3. Express the rare-entrance budget as a nonvisitation horizon and compare it with the sufficient training length for uniform restartability (Propositions 1 and 3).

Main proof of Remark 3.

Proof. With $N = N_\ell = |E_\ell|$ fixed and the constant choice in the proof of Proposition 1, rearranging Equation 21 gives

$$M' = \Omega\left(\frac{1}{KM\epsilon_1}\epsilon_2^{-N}\right), \qquad n = O\left(\frac{\delta M'}{B}\epsilon_2^{N}\right). \tag{126}$$

Thus the allowed nonvisit horizon grows linearly in M' with other quantities fixed. In addition to the first condition, if that horizon is ω of the sufficient learning scale in Equation 20, their ratio diverges. The comparison uses common bounds on the other problem parameters and confidence $1-\delta$ in each setting at each fixed chosen horizon. □

Remark 4 (The scale of initial rarity)**.** *The sufficient conditions of Proposition 1 imply $M' = \Omega(M^2 \log M)$, so M' is perfectly separated from M.*

Sketch of proof of Remark 4. Derive the minimum batch budget imposed by the learning condition and substitute it into the rarity requirement to obtain the coarse rarity scale (Lemma 29).

Lemma 29 (The coarse rarity scale)**.** *The sufficient conditions Equation 20 and Equation 21 imply $M' = \Omega(M^2 \log M)$ as $M \to \infty$.*

Proof. Since $\epsilon_1 < 1$, Equation 120 forces the drift to be positive. Squaring its defining strict inequality gives

$$\underline{\rho}^2_{\ell,K,M}\left(\sum_{j=1}^{n}\eta_j\right)^2 > \frac{\log(2KM/\delta)}{2B}\sum_{j=1}^{n}\eta_j^2.$$

Cauchy–Schwarz, $(\sum_j \eta_j)^2 \le n\sum_j \eta_j^2$, gives

$$Bn > \frac{\log(2KM/\delta)}{2\underline{\rho}^2_{\ell,K,M}}. \tag{127}$$

Substitution in Equation 21 gives

$$M' = \Omega\left(\left[\epsilon_2^{-N}\right]\frac{\log(2KM/\delta)}{\delta\underline{\rho}^2_{\ell,K,M}}\right). \tag{128}$$

Since $\underline{\rho}_{\ell,K,M} \le c_{\mathrm{route}}/M$ and $\epsilon_2^{-N} \ge 1$, for $M \to \infty$ this implies

$$M' = \Omega(M^2 \log M), \qquad M'/M \to \infty. \tag{129}$$

The fuller budget Equation 21 remains required for Proposition 1. □

Main proof of Remark 4.

Proof. Lemma 29 gives the claimed scale $M' = \Omega(M^2 \log M)$. This coarse implication does not replace the full budget Equation 21 required for Proposition 1. □

Remark 5 (Correct paths bypassing the rare entrance)**.** *On the event in the proof of Proposition 1, Equations 121 and 125 bound the final visitation probability of $u^* = u_m$ by $KM\epsilon_1$. Subtracting this probability from the original-start accuracy bound proved above gives*

$$\pi_{\mathbf{W}^{(n)}}\Big(y_T \in \tilde{C}_\ell,\ y_t \neq u^*\ \forall\, 0 \le t < T \,\Big|\, y_{\ell,0}, x_\ell\Big) \ge 1 - (3 + 2C_{\mathrm{bad}})KM\epsilon_1.$$

For $(3+2C_{\mathrm{bad}})KM\epsilon_1 < 1$, this probability is positive, so a correct path avoiding u^* exists. By the unique-entrance condition in Assumption 3, such a path also bypasses C_k.

C NUMERICAL EXPERIMENT

To ensure reproducibility, we provide the code for the numerical experiment in the zip file attached to the submission. The objective of this experiment is to compare finite-mini-batch training with the retention guarantees and population references associated with Theorems 3 and 4. Concrete numerical settings are listed below.

C.1 EXPERIMENT SETTINGS

Graph construction and initialization We instantiate the layered DAG constructed in Appendix B.1. The graph parameters and the choices used to construct $\mathbf{W}^{(0)}$ are summarized separately in Tables 1 and 2. All tasks use the same graph and all seeds use the same deterministic initialization.

Table 1: Graph parameters.

Parameter	Value
Tasks L	4
Clusters K	13
Node layers $(H+1)$	5
Nodes $\lvert V\rvert$	123
Directed edges $\lvert E\rvert$	214
Correct paths $\lvert\mathcal{P}_\ell\rvert$	10 per task
Maximum valid boundaries M	10
Boundary nodes $\sum_k \lvert\partial C_k\rvert$	53

Table 2: Initialization parameters.

Parameter	Value
Routing error parameter ϵ_1	0.02
Inter-cluster probability ϵ_2	0.04
Input dimension d	36
Task-specific input coordinates	4-D one-hot
Input squared norm $\lVert x_\ell\rVert_2^2$	33
Residual successors $\lvert R_k(u)\rvert$	1
Each start-to-entry probability	$1/10$
Boundary residual mass	0.92 or 0.88

At a task start, $|\partial^\star_\ell C_k| = 10$ and $I_{\ell,k}(s) = \varnothing$, so Equation 29 assigns probability $1/10$ to each entry boundary. At a relevant shared router, $|\partial^\star_\ell C_k| = 1$ and the remaining successors form $I_{\ell,k}(s)$; the correct exit receives probability $1/(1+\epsilon_1)$ and each other exit receives $\epsilon_1/((1+\epsilon_1)|I_{\ell,k}(s)|)$. For task inputs not using that router, its outgoing distribution is uniform. At every boundary, Equation 33 assigns probability ϵ_2 to each inter-cluster edge and probability $1 - d^{\mathrm{out}}_{u,k}\epsilon_2$ to its unique same-cluster successor. There are 51 boundaries with inter-cluster outdegree 2 and 2 with outdegree 3. With $K = 13$ and $\epsilon_2 = 0.04$, the sufficient condition in Appendix B.1 holds since $(K-1)\epsilon_2 = 0.48 < 1/2$.

The input matrix is $X = [\mathbf{1}_{4\times 32}\ \ I_4]$, where the first block has all entries equal to one. Thus $XX^\top$ has diagonal entries 33 and off-diagonal entries 32 and is invertible. We use Equation 35 directly:

$$\mathbf{W}^{(0)} e_{u,v} = X^\top (XX^\top)^{-1} \ell_{u,v}, \qquad \ell_{u,v} = \big(\log P_1(v \mid u), \ldots, \log P_4(v \mid u)\big)^\top. \tag{130}$$

The hint parameter $\epsilon_{\mathrm{rel}} = 0.1$ is used during SD training, not to construct $\mathbf{W}^{(0)}$.

Pairwise and cumulative conflict rates. Tables 3 and 4 report conflict rates, with rows denoting evaluated old tasks and columns denoting the tasks being learned. The pairwise rate θ compares only the row task and the column task. The cumulative rate θ_{total} is the fraction of the row task's correct paths that conflict with at least one task learned after it, up to and including the column task. Thus, θ_{total} is computed from a union of conflict-path sets: overlapping paths are counted only once, and the pairwise rates are not added. Diagonal entries are set to zero; entries below the diagonal are omitted.

Table 3: Pairwise conflict rate θ.

Old / New	Task 1	Task 2	Task 3	Task 4
Task 1	0	0.4	0.6	0.6
Task 2		0	0.9	0.9
Task 3			0	0.7
Task 4				0

Table 4: Cumulative conflict rate θ_{total}.

Old / New	Task 1	Task 2	Task 3	Task 4
Task 1	0	0.4	0.7	1.0
Task 2		0	0.9	0.9
Task 3			0	0.7
Task 4				0

For example, Task 1 has pairwise conflict rates 0.4 and 0.6 with Tasks 2 and 3, respectively, whereas their cumulative conflict rate is 0.7. In the plots, the term labeled "Limit" uses $\mathfrak{C}_{\mathrm{old}}(1-\theta_{\mathrm{total}})$, with

the concentration profile from Equation 89 constructed from the old task's correct-path probabilities at the end of its own training. This profile is held fixed throughout subsequent-task training, while θ_{total} is updated to reflect the cumulative conflict. Thus the plotted term uses the cumulative rate, rather than the pairwise θ, and is computed separately for each seed before averaging. The label refers to this concentration-profile term, not necessarily to the exact limiting correct generation probability.

Training and evaluation We train four tasks sequentially with 2×10^6 gradient updates per task, using mini-batches of size $B = 32$ and seeds $0, \ldots, 9$ for each method. The step size is $\eta_j = 11^{-1} j^{-0.01}$, where j starts at 1 for each task. The factor $11^{-1} = 3/\|x_\ell\|_2^2$ controls the update magnitude, while $j^{-0.01}$ provides a slow decay over successive updates. Plots show the mean correct generation probability and the minimum–maximum range across the ten seeds. Each probability is evaluated by summing the probabilities of all ten correct paths. For SD, the graph-specific population learning lower bound is shown during own-task training. At its end, the plot switches to the conditional retention lower bound anchored at the observed endpoint for each seed. The population learning curve is evaluated at $B \to \infty$ for the stability.

Graph-specific evaluation of constants. We tighten intermediate bounds in the Appendix proofs using the experimental graph, without altering the theorem statements. For Theorem 3, we use edge-specific hint and visitation probabilities in the odds bound and sum correct-path probabilities instead of applying the coarse coupling union bound. For Theorem 4, we evaluate the source-wise optimization and cross-task probability-transfer bounds using the actual visitation frequencies, logit offsets, and unchanged path factors. These refinements retain the underlying inequalities while avoiding worst-case estimates over the general graph family. The SD learning and SFT upper reference curves are evaluated in the population limit $B \to \infty$. The SFT conditional lower and SD retention bounds are instead evaluated at the observed phase endpoints, separately for each seed.

C.2 Experiment results

Table 5 reports the correct-generation probabilities after all four tasks have been learned. SFT reaches a mean probability above 0.999997 at the end of each task's own training, but its final means on Tasks 1–3 fall to 0.002799, 0.100312, and 0.304232, respectively. SD attains approximately 0.980386 on every task and preserves its own-training endpoint through the final task for every seed, to stored numerical precision. The final seed-to-seed range for SD is below 1.2×10^{-8} for each task.

Table 5: Final correct-generation probabilities after all four tasks. Entries are means [minimum, maximum] over ten seeds, rounded to six decimal places.

Task	SFT	SD
Task 1	0.002799 [0.002701, 0.002955]	0.980386 [0.980386, 0.980386]
Task 2	0.100312 [0.065468, 0.149365]	0.980386 [0.980386, 0.980386]
Task 3	0.304232 [0.263488, 0.334231]	0.980386 [0.980386, 0.980386]
Task 4	0.999998 [0.999998, 0.999998]	0.980386 [0.980386, 0.980386]

All saved evaluations satisfy the applicable SFT conditional lower and SD retention bounds, within numerical tolerance. The SFT trajectories also remain below the population upper references at the saved evaluations. For SD, 12 of 12,400 saved learning evaluations lie below the population learning reference, with a maximum shortfall of 7.83×10^{-6}. This distinction is consistent with using $B = 32$ for training: the $B \to \infty$ learning reference does not guarantee finite-mini-batch accuracy.

The following graphs are enlarged versions of the plots in Figure 2.

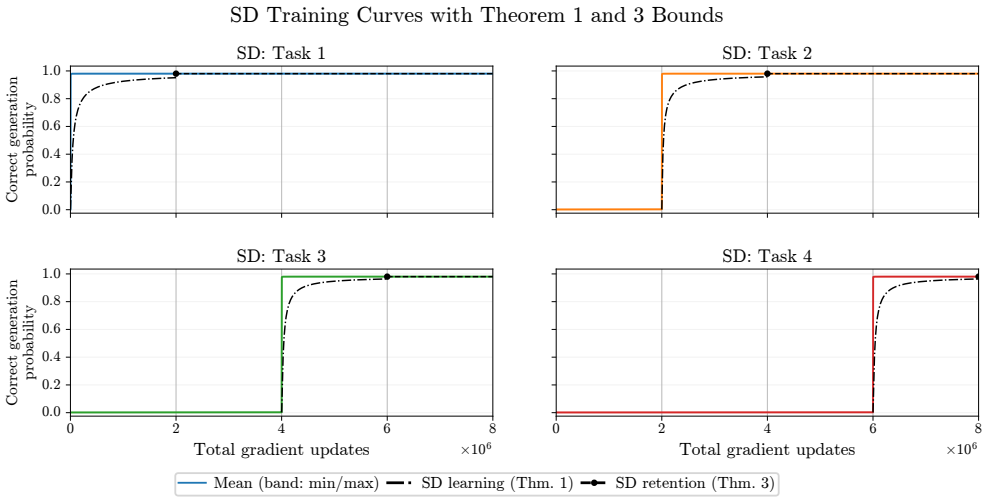

Figure 3: Training dynamics under SD for Tasks 1–4. The population learning lower reference is plotted during own-task training. At its end, the conditional retention lower bound is anchored at the observed endpoint and applies throughout subsequent-task training.

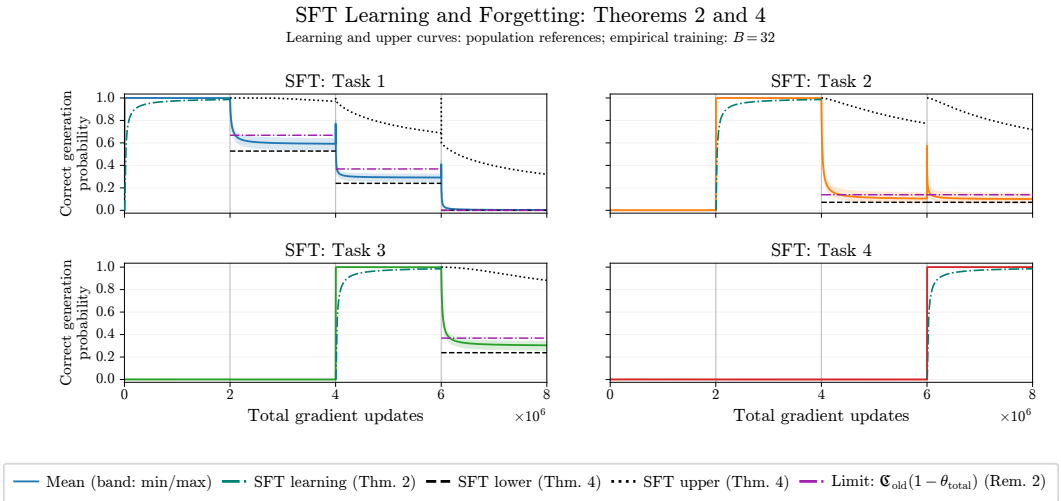

Figure 4: Training dynamics under SFT for Tasks 1–4. Black curves show conditional retention lower bounds and upper references. The concentration-profile term $\mathfrak{C}_{\mathrm{old}}(1-\theta_{\mathrm{total}})$ reflects the cumulative conflict and, for Task 1, the accumulation of forgetting across subsequent tasks.