BEYOND ANSWER QUALITY: PREDICTION AND EXECUTION UNDER PRIVILEGED DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent on-policy self-distillation (OPSD) methods have shown promise for improving language-model reasoning. A common form of OPSD uses privileged distillation, in which the teacher receives additional information, such as the correct answer, that is unavailable to the student. Yet studies of reasoning models also report cases where privileged supervision lowers answer accuracy and reduces the benefits of longer reasoning. Despite these observations, theoretical understanding of why such failures occur remains limited, making it difficult to understand when privileged supervision helps or hurts and how it should be improved. To address this gap, we develop a theoretical framework that separates two effects of privileged supervision: how well the student answers after completing the required reasoning steps (*prediction*), and whether it carries those reasoning steps through to completion (*execution*). We show that privileged supervision can make the student better at predicting the final answer once the required reasoning has been completed, while making it less likely to complete that reasoning on its own. As a result, better final-answer prediction can coexist with worse overall performance. We further show that this separation can emerge during finite distillation because learning at later reasoning steps depends on whether the student actually reaches those steps. Finally, we show a complementary effect on the teacher side: when privileged teachers are more likely to answer early, later reasoning steps are increasingly shaped by the teachers that continue reasoning. This changes what the student learns at those later steps, so different forms of privileged distillation can lead to different reasoning behavior. Experiments with Qwen3 and natural reasoning tasks provide qualitative support for these mechanisms.

1 INTRODUCTION

Recent on-policy self-distillation (OPSD) has emerged as a promising way to improve reasoning in language models by using a teacher copy of the model to supervise trajectories generated during training (Zhao et al., 2026). A common form of OPSD makes the teacher stronger by giving it information unavailable to the student, such as a reference answer or worked solution. This is a form of privileged distillation: the teacher can use the extra information when producing supervision, while the student must reason without it at inference time. Related privileged or demonstration-conditioned supervision has also been explored for reasoning, continual learning, and multi-turn agentic learning (Ho et al., 2023; Hsieh et al., 2023; Shenfeld et al., 2026; Penaloza et al., 2026).

These approaches have reported gains in reasoning and the acquisition of new capabilities (Zhao et al., 2026; Shenfeld et al., 2026). These gains, however, are not universal. Studies of thinking models report cases where privileged-context distillation reduces reasoning accuracy, shortens reasoning traces, weakens the benefit of larger generation budgets, or changes exploratory and self-corrective behavior (Kaur et al., 2026; Peng et al., 2026). Such effects matter because intermediate tokens can carry part of the reasoning itself rather than merely explain a final answer (Nye et al., 2021; Wei et al., 2022; Feng et al., 2023; Merrill & Sabharwal, 2024). Despite these observations, we still lack a clear theoretical account of why privileged supervision can help in some cases and hurt in others.

To understand this failure mode, we model privileged distillation by separating two effects that are usually entangled in next-token supervision. The first is **prediction**: how accurately the student can answer once the required reasoning steps have been completed. The second is **execution**: whether

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

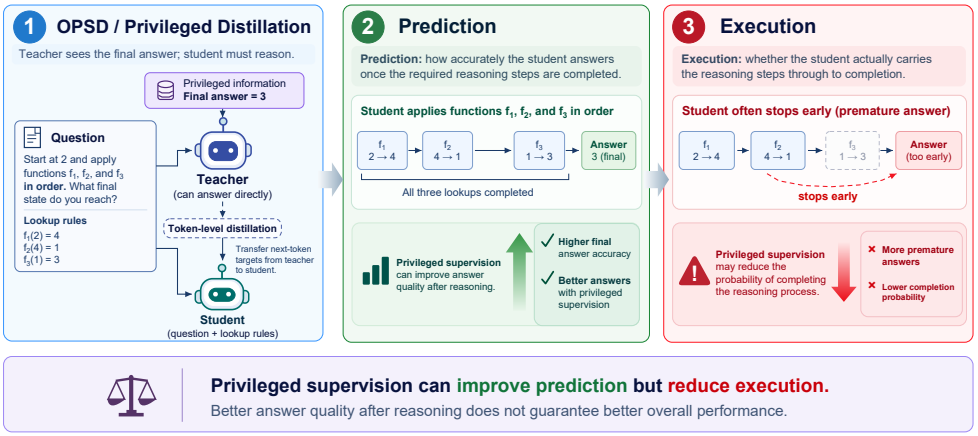


Figure 1: Privileged supervision can improve prediction while making the student less likely to complete the required reasoning steps.

the student actually carries those reasoning steps through to completion. This distinction reflects a simple asymmetry between teacher and student. A teacher that already sees the answer may reasonably stop early, while a student without that information may still need to continue reasoning. Because token-level distillation matches the teacher's full next-token distribution, it can transfer both the teacher's answer signal and its tendency to stop. Our framework makes these two effects separately analyzable; Figure 1 summarizes this asymmetry.

Using this framework, we first show that privileged supervision can improve final-answer prediction while making the student less likely to complete the reasoning needed to reach that answer, so better prediction can coexist with worse overall performance. This perspective raises two further questions. How can the prediction–execution separation emerge during distillation, rather than simply being imposed by a particular final policy? And when a privileged teacher is more likely to answer early, how does this reshape the supervision available at later reasoning steps?

We answer these questions by studying a step-by-step reasoning task in which the final answer is obtained through a sequence of intermediate reasoning steps, while the teacher may additionally observe the correct answer. Within this framework, we characterize when better final-answer prediction can coincide with lower completion, how this separation can emerge during finite distillation, and how early answers from privileged teachers reshape the supervision available at later reasoning steps. Our contributions are as follows.

1. **A new prediction–execution perspective.** We separate how well the student predicts the final answer after the required reasoning has been completed from whether it completes that reasoning in the first place. Within this perspective, we show conditions under which privileged supervision improves final-answer prediction but lowers completion enough that overall performance falls (Theorem 1).
2. **The separation can emerge during finite distillation.** We show that later reasoning steps receive less effective training when the student reaches them less often, and prove that the prediction–execution separation can emerge after finitely many updates (Theorem 2).
3. **Early teacher answers reshape later supervision.** When privileged teachers are more likely to answer early, they are less represented at later reasoning steps. We show that even when the same distribution of later-step reasoning contexts is observed, the student can learn different behavior depending on whether supervision remains tied to the teacher that reached those steps or is assigned independently (Theorem 3). Experiments with Qwen3 and natural reasoning tasks provide qualitative support for these mechanisms.

2 Related Work

2.1 Privileged Information and On-Policy Distillation

Knowledge distillation transfers predictive information from a teacher to a student and has been extended from output distributions to sequence-level supervision (Hinton et al., 2015; Kim & Rush,

2016). Learning using privileged information studies a complementary setting in which additional information is available during training but unavailable at test time, and generalized distillation connects these two perspectives (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2016). In sequential prediction, the states on which supervision is provided also depend on the rollout policy: DAgger trains on learner-induced states, while scheduled sampling mixes observed and generated histories (Ross et al., 2011; Bengio et al., 2015). Recent language-model distillation methods increasingly exploit this interaction. GKD performs distillation on student-generated sequences, while MiniLLM and DistiLLM develop related objectives and sampling strategies for language-model distillation (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). More recent self-distillation methods condition the teacher on additional context or demonstrations unavailable to the student, with applications to mathematical reasoning, continual learning, skill and knowledge acquisition, and multi-turn agentic learning (Zhao et al., 2026; Shenfeld et al., 2026; Penaloza et al., 2026). Distillation has also long been used to preserve previously acquired capabilities in continual and incremental learning (Li & Hoiem, 2016; Rebuffi et al., 2017). Recent analyses further show that on-policy self-distillation may alter reasoning behavior or reduce output diversity (Kaur et al., 2026; Peng et al., 2026; Nicolicioiu et al., 2026). Our analysis separates prediction after completed computation from the probability of completing it, and studies both finite student updates and how teacher-generated contexts acquire their supervision.

2.2 Reasoning Supervision and Intermediate Computation

Chain-of-thought and scratchpad methods make intermediate generation an explicit component of language-model reasoning (Nye et al., 2021; Wei et al., 2022). A large body of work subsequently uses generated rationales or intermediate supervision to transfer reasoning capabilities: Fine-tune-CoT learns from teacher-generated reasoning traces, Socratic CoT distills problem decompositions and intermediate steps, Distilling Step-by-Step uses rationales as auxiliary supervision, and STaR iteratively generates and learns from reasoning trajectories (Ho et al., 2023; Shridhar et al., 2023; Hsieh et al., 2023; Zelikman et al., 2022). Process supervision instead evaluates intermediate reasoning steps directly (Lightman et al., 2024). Beyond serving as explanations or training targets, intermediate tokens can themselves provide computational resources. Constructive analyses show that chain of thought can increase the capability of specified Transformer architectures on multistep tasks (Feng et al., 2023; Merrill & Sabharwal, 2024). At inference time, related work studies how additional reasoning computation should be used or allocated, including self-consistency over multiple reasoning paths, adaptive allocation of test-time compute, and explicit budget forcing (Wang et al., 2023; Snell et al., 2024; Muennighoff et al., 2025). In contrast, we study how distillation changes the execution of intermediate computation itself.

3 Problem Setup

We use a step-by-step lookup task to make **prediction** and **execution** separately observable. A student that is not given the final answer must carry out a sequence of intermediate computations, whereas a teacher that already sees the answer may have no need to complete the same sequence. Precise definitions and network constructions are collected in Appendix B.

A task that requires sequential computation. The input contains a starting state and a sequence of lookup tables. Applying one table produces the state used by the next. For example, a three-table task might follow

$$2 \xrightarrow{f_1} 4 \xrightarrow{f_2} 1 \xrightarrow{f_3} 3.$$

A full execution reports 4 and 1, then answers 3. A student that is not given the answer must carry the state through these lookups to discover 3, while a teacher that already knows the answer can respond immediately. This is the information asymmetry we study.

Formally, there are $K \geq 2$ tables and $M \geq 2$ states, written $[M] = \{1, \ldots, M\}$. Each f_j is an independently and uniformly drawn one-to-one mapping, and S_0 is uniform and independent of the tables. Fresh tables and a fresh start are drawn for every task. The computation and answer are

$$S_j = f_j(S_{j-1}) \quad (j = 1, \ldots, K), \qquad Y = S_K. \tag{3.1}$$

The complete input $X = (f_1, \ldots, f_K, S_0)$ determines Y. There are $d = K - 1$ intermediate-state outputs, and $u = 1/M$ denotes uniform-guess accuracy.

What the student can do. At an intermediate step, the student either reports a stage and state and continues, or emits an answer and stops. We write $\mathcal{C}$ and $\mathcal{A}$ for these disjoint token sets. Reporting stage t and state s makes the next lookup $f_{t+1}(s)$, even if the report is wrong or out of order; generated states are never replaced with correct ones. On a correct path, $t = d$ is the terminal stage, where the last lookup yields Y.

The student sees the full task input, but its trainable readout uses only the local feature produced by the current lookup. Thus it cannot obtain the final answer by directly reading through all remaining tables; it must carry the intermediate state through the computation. We implement this by freezing attention and feature extraction and restricting the trainable output layer to the current lookup feature. The three results use different output parameterizations within this shared interface.

What the teacher knows. Teacher and student start from the same initial model p_0, and the teacher remains fixed. With probability ρ, the teacher additionally sees the correct answer Y; otherwise it receives an empty reference $\perp$. The student always receives $\perp$. We write q_0 and q_1 for the teacher's next-token distributions without and with the answer reference. The privilege choice is independent of the task and fixed within a trajectory. Section 4.3 later distinguishes whether supervision on a teacher-generated prefix keeps this choice or uses an independent one.

What distillation transfers. Next-token distillation does not teach only *what answer to produce*. At an unfinished step, the teacher distribution also specifies *whether to continue or answer now*. Matching the full distribution can therefore transfer both answer information and the teacher's tendency to stop.

A *prefix* is the sequence generated so far. At each prefix used for distillation, the student matches the teacher's full next-token distribution by forward KL; losses are summed over visited prediction positions, including the first answer position, without length normalization. At a fixed prefix, let q, p be teacher and student distributions and let $c_q = q(\mathcal{C})$, $c_p = p(\mathcal{C})$ be their continuation probabilities. Let $q_{\mathcal{C}}, p_{\mathcal{C}}$ denote the outputs conditional on continuing and $q_{\mathcal{A}}, p_{\mathcal{A}}$ those conditional on answering. When both branches have positive mass,

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) = {} & D_{\mathrm{KL}}(\mathrm{Ber}(c_q)\|\,\mathrm{Ber}(c_p)) \\ & + c_q D_{\mathrm{KL}}(q_{\mathcal{C}}\|p_{\mathcal{C}}) + (1-c_q) D_{\mathrm{KL}}(q_{\mathcal{A}}\|p_{\mathcal{A}}). \end{aligned} \tag{3.2}$$

The first term matches the continue-versus-answer decision; the other two match what is produced after that choice. Privileged distillation can therefore improve the answer signal while also teaching a stopping tendency that may not suit a student that still needs the intermediate computation. Boundary cases use the usual extended-KL convention. Each result below specifies how its prefixes and teacher targets are obtained.

Prediction, execution, and evaluation. We measure answer quality and completion separately. Free generation stops at the first answer or after K outputs; a correct first answer counts as success even after an incorrect continuation, while no answer within the cap counts as failure.

Quantity	What it asks
Free accuracy A_{free}	Can the student solve the task when it chooses when to continue and when to answer?
Completion P_{complete}	Does the student generate all d correct intermediate states before answering?
Forced accuracy A_{forced}	How well does the student perform when early answers are suppressed, without correcting its intermediate computation?

Forced evaluation supplies no correct intermediate states and repairs no errors. Theorem 1 suppresses answers until the terminal stage; because every continuation is correct in that setting, A_{forced} equals terminal prediction accuracy there. Theorem 2 suppresses answers for the first d outputs and leaves the final output unmasked. We do not identify forced accuracy with terminal prediction accuracy in general.

These definitions provide the common language for the three results below.

4 MAIN RESULTS

Section 3 separates final-answer prediction from execution. Our first result makes this *prediction–execution* perspective precise, showing that privileged supervision can improve prediction while reducing completion enough to hurt overall performance. This separation raises two complementary questions: on the student side, can finite distillation itself produce it; and on the teacher side, how do earlier privileged answers change the supervision available at later reasoning steps? Sections 4.2 and 4.3 answer these questions in separate analytical settings. Complete assumptions and proofs are in Appendix B onward.

4.1 A PREDICTION–EXECUTION SEPARATION

Our first result isolates the central tension. Before the computation is complete, the student's local readout does not yet identify the final answer. A privileged teacher may nevertheless put substantial probability on answering immediately because it already knows that answer. Across tasks, unfinished reasoning states do not reveal which final answer is correct, so the answer identities themselves average out while the teacher's tendency to answer early remains. Once the computation is complete, however, the same privileged supervision provides a better answer signal. The student can therefore inherit both a better terminal predictor and a stronger tendency to stop before reaching it.

We formalize this effect by minimizing expected forward KL with the prefix distribution held fixed. A population fit is *self-consistent* when the fitted student generates that same prefix distribution. In Setting 1, continuing is error-free, so execution is isolated entirely through the decision of whether to continue. The unreferenced teacher continues until terminal and then answers through channel Q_0, with mean accuracy $a_0 > u$, while the referenced teacher already knows the answer and answers correctly immediately. Here $Q_0(i \mid y)$ is the probability of answering i at a correct terminal state with answer y, and I_M denotes the perfect-answer channel.

Theorem 1 (Prediction and survival)**.** *Under Setting 1, for $0 \le \rho < 1$, self-consistent population fits exist and share the terminal channel $Q_\rho = (1-\rho)Q_0 + \rho I_M$. Writing $a_\rho = (1-\rho)a_0 + \rho$, their accuracies satisfy*

$$\boxed{A_{\text{free}} = u + (a_\rho - u)P_{\text{complete}}}, \qquad P_{\text{complete}} = (1-\rho)^d, \qquad A_{\text{forced}} = a_\rho. \tag{4.1}$$

Here P_{complete} is the probability of generating all d correct continuations. If $(1+du)/(d+1) < a_0 < 1$, then, with K, M, Q_0 fixed, A_{forced} strictly increases and A_{free} strictly decreases as ρ increases on $[0, 1)$.

Equation (4.1) makes the prediction–execution perspective explicit. Privilege improves the terminal answer channel a_ρ, but at the same time lowers the probability of reaching the point where that improved answer channel can be used. Overall performance depends on both. Under the stated condition, completion falls fast enough that the student becomes better at answering after continued execution while becoming worse at solving the task on its own.

For fixed M, the threshold $u+(1-u)/(d+1)$ decreases with computation depth, so completion loss more readily dominates the terminal gain. This does not imply that shorter reasoning is generally harmful: outside Setting 1 an early answer can itself be correct. In general $A = Pa + (1-P)b$ for completion probability P and conditional accuracies a, b; Theorem 1 isolates $a = a_\rho, b = u$, and forced accuracy need not equal terminal prediction accuracy outside this setting.

4.2 HOW THE SEPARATION CAN EMERGE DURING FINITE DISTILLATION

Theorem 1 describes a fitted student. We now turn to the first follow-up question: can finite distillation itself move the student toward the same prediction–execution separation? The key observation is simple. A later reasoning step receives a direct update only on trajectories that reach it. If the student becomes less likely to continue, later steps—including the terminal answer position—receive less effective training.

Setting 2 makes this dependence explicit through finitely many exact population updates of a softmax model with parameters $\theta = (z_1, \ldots, z_d, w)$: z_j controls continuation at depth j and w controls

terminal prediction, while the remaining output rows are suppressed by a finite $L > 0$. Let θ_n denote the actual network after n updates of size $0 < \eta \leq 1$, and write F_L, G_L for its free and forced accuracies.

The comparison process used for analysis has no-reference/reference continuation targets $0 < c_{\mathrm{ref}} < c_0 < 1$ and terminal accuracies $u < a_0 < a_{\mathrm{ref}} < 1$. Define $\Delta_a = a_{\mathrm{ref}} - a_0$, $\Delta_c = c_0 - c_{\mathrm{ref}}$, and the privilege-averaged targets $c_* = c_0 - \rho\Delta_c$, $a_* = a_0 + \rho\Delta_a$. The comparison process starts from the same logits as the actual network. With $\sigma(x) = (1 + e^{-x})^{-1}$, write $c_j(n) = \sigma(\bar{z}_j(n))$ and $a(n) = \sigma(\bar{w}(n))$. Appendix B.4 gives the full construction.

Theorem 2 (Finite dynamics and the terminal visit clock)**.** *Under Setting 2, with $0 < \rho < 1$, the following hold.*

(i) Comparison dynamics. *The comparison process satisfies*

$$\bar{z}_j(n+1) - \bar{z}_j(n) = \eta \prod_{i<j} c_i(n)\,[c_* - c_j(n)], \tag{4.2}$$

$$\bar{w}(n+1) - \bar{w}(n) = \eta \prod_{j=1}^{d} c_j(n)\,[a_* - a(n)]. \tag{4.3}$$

Consequently $c_j(n) \downarrow c_*$ and $a(n) \uparrow a_*$.

(ii) Terminal training exposure. *The terminal error obeys*

$$H_N = \eta \sum_{n=0}^{N-1} \prod_{j=1}^{d} c_j(n), \qquad 0 \leq a_* - a(N) \leq \rho\Delta_a e^{-\lambda_a H_N}, \tag{4.4}$$

where $\lambda_a = \min\{a_0(1-a_0), a_*(1-a_*)\}$.

(iii) Finite-network separation. *If*

$$\Delta_a c_0 < d\,\Delta_c(a_0 - u), \tag{4.5}$$

then $D = (a_0 - u)c_0^d - (a_* - u)c_*^d > 0$. For every sufficiently large finite N, a sufficiently large finite L, depending on N, gives the actual finite network

$$G_L(\theta_N) - G_L(\theta_0) \geq \tfrac{1}{2}\rho\Delta_a > 0, \qquad F_L(\theta_0) - F_L(\theta_N) \geq \tfrac{1}{2}D > 0. \tag{4.6}$$

The three parts describe one student-side mechanism. Updates at depth j are weighted by the probability of reaching that depth; in particular, H_N acts as an *effective terminal training clock*. Terminal prediction therefore continues to learn, but on a clock slowed by reduced execution. Under condition (4.5), finitely many updates suffice for the actual finite network to end with higher forced and lower free performance. Because $c_* > 0$, exposure slows without ceasing permanently; monotonicity applies to the comparison process, while the finite network inherits the endpoint guarantee in (4.6).

4.3 How Early Teacher Answers Reshape Later Supervision

We now turn to the second, parallel follow-up question. Section 4.2 studied the student side: how often a later step is reached controls how much training it receives. Here we study the teacher side: when privileged teachers answer earlier, what supervision remains associated with the reasoning paths that reach later steps?

Suppose teacher branch b has continuation probability c_b and terminal answer channel Q_b, with mean accuracy a_b. Setting 3 assumes $0 \leq c_1 < c_0 \leq 1$ and $u < a_0 \leq a_1 \leq 1$: the referenced teacher is at least as accurate at terminal but less likely to continue. Consequently, privileged teacher trajectories become progressively underrepresented among paths that survive to later reasoning steps. In the extreme case $c_0 = 1, c_1 = 0$, every deep teacher-generated prefix comes from the unreferenced branch.

Once a later context has been collected, its target can remain tied to the teacher condition that generated it (**RETAIN**) or use an independently chosen teacher condition (**RELABEL**). RETAIN

preserves the association between survival and privilege; RELABEL keeps the same collected-prefix distribution but breaks that association. We also analyze ON, where the student generates prefixes, and PROCESS, which uses unprivileged intermediate supervision and privileged mixing only at terminal.

Formally, the collection bit B_{col} and each newly drawn labeling bit B_{lab} have Bernoulli(ρ) marginals and are reused within a trajectory; independent labeling means independence of the task and collection.

Protocol	Prefix collection	Teacher target
RETAIN	Frozen teacher with B_{col}	$q_{B_{\mathrm{col}}}$: retain the collection bit.
RELABEL	Same teacher-collection law	$q_{B_{\mathrm{lab}}}$: draw an independent bit.
ON	Student	$q_{B_{\mathrm{lab}}}$: draw an independent bit.
PROCESS	Student	q_0 at $t < d$; $q_{B_{\mathrm{lab}}}$ at $t = d$.

PROCESS does not provide gold intermediate states; it is used only to separate intermediate from terminal privileged supervision.

Theorem 3 (Selection and conditional fitting)**.** *Under Setting 3, let* $0 < \rho < 1$ *and* $D_j = (1-\rho)c_0^j + \rho c_1^j$ *for* $0 \le j \le d$*, with* $c_b^0 = 1$.

(i) Survival selection. *Among teacher trajectories surviving j continuations, the privilege posterior is*

$$w_j = \Pr(B_{\mathrm{col}} = 1 \mid \textit{survive } j) = \frac{\rho c_1^j}{(1-\rho)c_0^j + \rho c_1^j}. \tag{4.7}$$

(ii) Conditional fits. *Conditional population fits exist. For $0 \le j < d$, their continuation probabilities are*

$$c_{\mathrm{RETAIN},j} = \frac{D_{j+1}}{D_j}, \qquad c_{\mathrm{RELABEL},j} = (1-\rho)c_0 + \rho c_1. \tag{4.8}$$

Their terminal channels are

$$Q_{\mathrm{RETAIN}} = (1-w_d)Q_0 + w_d Q_1, \qquad Q_{\mathrm{RELABEL}} = (1-\rho)Q_0 + \rho Q_1. \tag{4.9}$$

(iii) Protocol comparison. *ON has the same fitted continuation and terminal channel as RELABEL. In the source specialization $c_0 = 1$, $c_1 = 0$, $Q_1 = I_M$, the ON/RETAIN accuracy ordering depends on a_0 and K; for each fixed $K > 2$, $M \ge 2$, $0 < \rho < 1$, both strict orderings occur among admissible Q_0.*

The first part quantifies selection: because $c_1 < c_0$, the privilege mixture among surviving trajectories changes with depth. The second part shows why this matters. RETAIN and RELABEL can expose the student to the *same distribution of collected reasoning contexts* yet fit different policies because they attach different teacher distributions to those contexts. What matters is therefore not only which reasoning steps are observed, but how supervision remains associated with the process that produced them.

The final comparison rules out a universal protocol ranking: either ON or RETAIN can be more accurate depending on the answer channel. Viewed through the prediction–execution perspective, this lack of a universal ordering is natural: changing how privilege is assigned can affect both terminal prediction and the tendency to continue, so privilege alone does not determine a uniformly preferred protocol. PROCESS preserves the initial continuation target, $c_{\mathrm{PROCESS},j} = c_0$, while using RELABEL's terminal mixture, $Q_{\mathrm{PROCESS}} = (1-\rho)Q_0 + \rho Q_1$.

5 Experiments

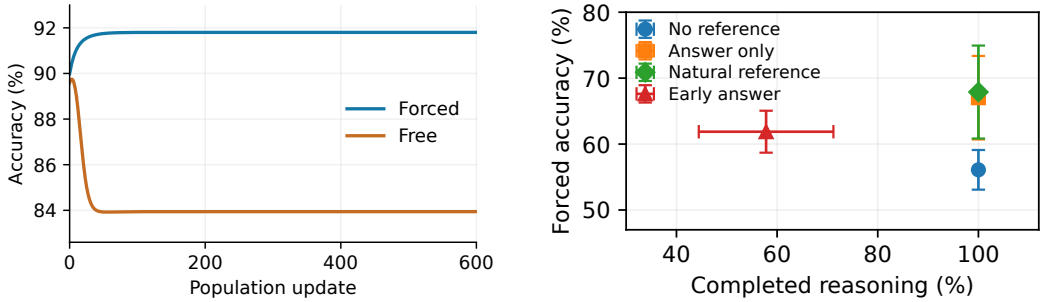
We examine how answer guidance and the decision to keep reasoning interact during learning, first numerically and then in Qwen3. Complementary studies examine student reasoning length on GSM8K and teacher continuation with and without a reference answer. Settings and additional results are in Appendix A.

5.1 Answer guidance and execution during learning

A numerical training trajectory. Over 600 population updates, forced accuracy rises from 90.00% to 91.80% while free accuracy falls from 89.60% to 83.94% (Figure 2a), illustrating Theorem 2's finite-training separation. Appendix A.2.3 details the run and sufficient-bound comparison.

Separating two kinds of teacher guidance. Qwen3-1.7B (Qwen Team, 2025) performs five sequentially revealed lookups. The environment then applies a hidden final lookup to the student's actual state and reveals its result for answer mapping; wrong states are never repaired. We train LoRA (Hu et al., 2021) with full-vocabulary token-level forward KL and sample legal actions. Forced evaluation suppresses early answers without supplying intermediate states.

We compare four forms of guidance. Reference use is selected independently per trajectory with probability $1/2$. **No reference (BASE)** always uses the unreferenced teacher. When reference use is selected, **Answer only (TERM-ONLY)** uses it only for the final answer, and **Natural reference (NATURAL-FULL)** uses it throughout. **Early answer (EARLY-PRIV)** keeps TERM-ONLY's terminal target but uses an 80% continue/20% correct-early-answer mixture at intermediate decisions, preserving relative probabilities of complete legal continuations. Otherwise all targets remain unreferenced. All settings share the same unprivileged initialization, five paired fitting seeds, and 256 test tasks; update 8 is the prespecified primary endpoint. The adapter also changes internal token targets; Appendix A.1.3 specifies this intervention.



(a) A numerical training trajectory

(b) Teacher guidance in Qwen3

Figure 2: **Better answers need not mean more completed reasoning. (a)** Forced and free accuracy separate during numerical training. **(b)** Qwen3 at update 8: terminal guidance improves forced accuracy; early-answer guidance reduces completion. Points and bars are five-fit means and sample standard deviations.

Table 1: Qwen3 at update 8: five-fit means (%). Individual seeds are in Appendix A.1.4.

Teacher guidance	Forced accuracy	Completion	Free accuracy	Premature answers
No reference	56.09	100.00	56.25	0.00
Answer only	67.03	100.00	66.56	0.00
Natural reference	67.89	100.00	68.20	0.00
Early answer	61.88	57.81	41.48	42.11

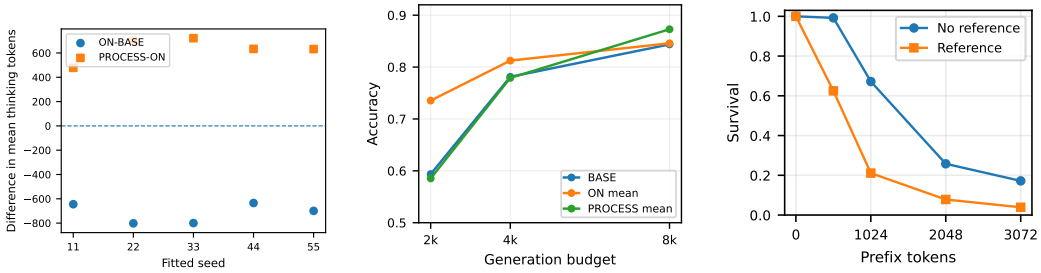
Answer-only guidance improves prediction. Answer-only guidance gains 10.94 points of forced accuracy over no-reference guidance in the five-seed mean, with every seed improving and completion unchanged at 100%. Natural reference guidance also retains full completion; the teacher audit finds continuation probability almost one at intermediate steps (Table 1).

Early-answer guidance reduces completion. Early-answer guidance lowers completion by 42.19 points relative to answer-only guidance. Of 540 first execution failures across five runs, 539 are premature answers and one an incorrect continuation; correctness among emitted continuations averages 99.98%.

Relative to matched no-reference guidance at update 8, early-answer guidance has higher forced accuracy ($+5.78$ points), lower completion (-42.19), and lower free accuracy (-14.77), with all three directions holding in every seed. Thus the student performs better when required to continue, but worse when it chooses when to answer. Full curves are in Appendix A.1.5.

5.2 Reasoning length on GSM8K

Does answer accuracy capture all changes in reasoning? We fit Qwen3-1.7B with ON and PROCESS for 32 updates, using reference probability $1/2$ and five seeds. PROCESS keeps reference-based guidance out of the reasoning phase. All models answer the same 96 GSM8K questions (Cobbe et al., 2021); prefixes of each 8192-token stream supply shorter-budget evaluations.



(a) Student reasoning length **(b)** Accuracy at each budget **(c)** Teachers still reasoning

Figure 3: **Student reasoning and teacher continuation on real questions.** **(a)** ON uses fewer thinking tokens than BASE in all five seeds; PROCESS uses more than ON. **(b)** Accuracy on prefixes of the same generated stream, with five-seed means for ON and PROCESS. **(c)** Teachers given a reference answer are less represented among trajectories that continue to later positions.

ON uses 635–802 fewer thinking tokens than BASE in every seed; PROCESS restores 478–722 relative to ON (Figure 3a). ON's 8192-token accuracy remains close to BASE, yet it is shorter on 87–92% of questions both answer correctly. Answer accuracy alone misses this consistent change in generated reasoning length, which is compatible with more concise reasoning. The number of thinking tokens measures generated reasoning length, not required-step completion.

5.3 How far teachers continue reasoning

On the teacher side, we pair referenced and unreferenced Qwen3 trajectories on 128 questions to examine who still contributes at later steps. At 1024 tokens, 86/128 unreferenced trajectories are still reasoning (67.2%), compared with 27/128 referenced trajectories (21.1%). The referenced share among those remaining falls from 50% to 23.9% (Figure 3c). This illustrates the selection effect in Theorem 3. RETAIN/RELABEL comparisons show no stable accuracy or length ordering (Appendix A.2.5).

6 Conclusion and Limitations

Privileged distillation transfers not only answer information but also decisions about whether computation should continue. Our prediction–execution perspective shows that better final-answer prediction can coexist with lower completion and worse overall performance. It also motivates two complementary analyses: finite distillation couples later learning to how often the student reaches later steps, while early privileged answers reshape the supervision available at those steps and can change the fitted policy. Qwen3 interventions and natural-task studies qualitatively support these mechanisms. Our guarantees rely on the specified readouts, population objectives, and data-generation assumptions. Overall, privileged distillation should be analyzed through prediction, execution, and where supervision enters reasoning, not answer quality alone.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *ICLR*, 2024. URL https://arxiv.org/abs/2306.13649v3.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks, 2015. URL https://arxiv.org/abs/1506.03099v3. arXiv preprint arXiv:1506.03099.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168. arXiv preprint arXiv:2110.14168.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.15408v5.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *ICLR*, 2024. URL https://arxiv.org/abs/2306.08543v4.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015. URL https://arxiv.org/abs/1503.02531. arXiv preprint arXiv:1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large Language Models Are Reasoning Teachers. In *ACL*, pp. 14852–14882, 2023. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of ACL*, 2023. URL https://aclanthology.org/2023.findings-acl.507/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. URL https://arxiv.org/abs/2106.09685. arXiv preprint arXiv:2106.09685.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models, 2026. URL https://arxiv.org/abs/2607.05184v1. arXiv preprint arXiv:2607.05184.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *EMNLP*, pp. 1317–1327, 2016. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629, 2016. doi: 10.1007/978-3-319-46493-0_37. URL https://zhizhongli.vision/projects/learning-without-forgetting/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *ICLR*, 2024. URL https://arxiv.org/abs/2305.20050.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying Distillation and Privileged Information. In *ICLR*, 2016. URL https://arxiv.org/abs/1511.03643v3.

William Merrill and Ashish Sabharwal. The Expressive Power of Transformers with Chain of Thought. In *ICLR*, 2024. URL https://arxiv.org/abs/2310.07923.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple Test-Time Scaling, 2025. URL https://arxiv.org/abs/2501.19393v3. arXiv preprint arXiv:2501.19393.

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity, 2026. URL https://arxiv.org/abs/2606.26091v1.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show Your Work: Scratchpads for Intermediate Computation with Language Models, 2021. URL https://arxiv.org/abs/2112.00114. arXiv preprint arXiv:2112.00114.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models, 2026. URL https://arxiv.org/abs/2602.04942.

Keqin Peng, Chen Li, Yuanxin Ouyang, Yancheng Yuan, and Liang Ding. Diagnosing and Mitigating Thinking Collapse in On-Policy Self-Distillation, 2026. URL https://arxiv.org/abs/2607.10805v1. arXiv preprint arXiv:2607.10805.

Qwen Team. Qwen3-1.7B model card and release, 2025. URL https://huggingface.co/Qwen/Qwen3-1.7B.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Rebuffi_iCaRL_Incremental_Classifier_CVPR_2017_paper.html.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *AISTATS*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897v2.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling Reasoning Capabilities into Smaller Language Models. In *Findings of ACL*, pp. 7059–7073, 2023. URL https://aclanthology.org/2023.findings-acl.441/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters, 2024. URL https://arxiv.org/abs/2408.03314. arXiv preprint arXiv:2408.03314.

Vladimir Vapnik and Rauf Izmailov. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL https://www.jmlr.org/papers/v16/vapnik15b.html.

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 *ICLR*, 2023. URL https://arxiv.org/abs/2203.11171v4.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2201.11903.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2203.14465v2.

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, 2026. URL https://arxiv.org/abs/2601.18734v3. arXiv preprint arXiv:2601.18734.

APPENDIX

AI USE STATEMENT

The human authors determined the research topic and problem formulation. We used AI to assist with manuscript drafting and polishing, developing and checking mathematical arguments, experiment design and implementation. The authors take responsibility for the final text, mathematical claims, experimental results, and associated artifacts.

Contents

A Supplementary experiments 13
A.1 Qwen3 teacher guidance and execution 13
A.2 Earlier experiments and extended results 18
A.3 Additional empirical details 24

B Theoretical settings and assumptions 25
B.1 Task, evaluation, and fitting definitions 25
B.2 Compact theorem-level settings 26
B.3 Theorem 1 setting 27
B.4 Theorem 2 setting 30
B.5 Explicit finite-horizon bounds 33
B.6 Theorem 3 setting 33

C Proofs of the main theorems 36
C.1 Theorem 1 36
C.2 Theorem 2 38
C.3 Theorem 3 39

D Supporting lemmas and propositions 41
D.1 Theorem 1 41
D.2 Theorem 2 52
D.3 Theorem 3 66

A SUPPLEMENTARY EXPERIMENTS

A.1 QWEN3: ANSWER GUIDANCE AND THE DECISION TO CONTINUE

This study compares four teacher-target policies on the same five-step task, starting from a shared Qwen3-1.7B initialization. The primary comparison is at update 8, fixed before the intervention runs. The four arms, five fitting seeds, and six evaluated checkpoints give 20 fits and 120 evaluation cells. This subsection documents the new experiment in Section 5.1; the earlier full-context lookup experiment is retained separately in Appendix A.2.2.

A.1.1 TASK, INFORMATION ACCESS, AND INITIALIZATION

Each instance contains six independent random permutations $f_1, \ldots, f_6$ of eight states and a starting state S_0. The intended path is $S_j = f_j(S_{j-1})$, and the correct answer is $Y = g(S_6)$ for a fixed codebook g. The student emits five continuation actions. At intermediate decision depth $j = 0, \ldots, 4$, it sees its actual current state, f_{j+1}, and its own history. Future tables, the final answer, the codebook mapping, and an identifying task ID are absent from the prompt. A reported state becomes the next current state, including when it is wrong.

After five continuations, the environment computes $\widehat{S}_{\mathrm{term}} = f_6(\widehat{S}_5)$ from the student's *actual* reported state and reveals only $\widehat{S}_{\mathrm{term}}$; f_6 itself remains hidden. The student then predicts $g(\widehat{S}_{\mathrm{term}})$. Scoring always uses $g(f_6(S_5))$ from the true task. Thus the terminal feature supplies the result of the last lookup, but does not repair the preceding computation. A counterfactual audit holds each visible intermediate prompt fixed while varying hidden information: all 960 checks passed, with different final answers still possible. This is an information-access bottleneck at the task interface; the theorem instead specifies a restricted readout with the full task input present.

The fixed codebook is

s	0	1	2	3	4	5	6	7
$g(s)$	Q	X	W	L	B	J	R	V

and is never printed for the student. Labels were chosen from tokenization only, before model outcomes. Each label is one token in isolation; each complete `ANSWER(label)` string uses four tokens. The actual action tokenization is used for scoring and teacher construction, not the isolated label IDs.

We use Qwen3-1.7B revision `70d244cc86ccca08cf5af4e1e306ecf908b1ad5e` with LoRA rank 8, scaling 16, and dropout 0. Calibration uses no-reference supervised cross-entropy, starting from an unprivileged computation checkpoint. All 33 even checkpoints from 64 through 128 were evaluated in increasing order on a separate 128-task selection set. Update 98 was the first to satisfy forced accuracy in $[0.55, 0.80]$, completion at least 0.90, and correctness given continuation at least 0.95. Its selection-set values were 57.81%, 100%, and 100%, respectively. No referenced-teacher calls or formal treatment results were used in selection. Earlier unsuccessful task constructions and the dense-selection record remain archived; the dense run was a new calibration study, not a replay of the earlier trajectory.

The fixed formal test set contains 256 tasks. On this set, the shared pre-distillation checkpoint has forced and free accuracy 62.89% and completion 100%. The main result compares *matched arms after eight updates*, not improvement of EARLY-PRIV over its own initial checkpoint. In particular, EARLY-PRIV's update-8 forced mean is 61.88%, compared with the matched BASE mean of 56.09%.

A.1.2 TEACHER GUIDANCE, FITTING, AND EVALUATION

Write q_0 and q_1 for the frozen teacher without and with the true coded answer. The referenced prompt replaces only the reference-answer field; it retains the student's actual state and history. Every arm uses the same initialization, task order, and paired random-number streams where applicable. A Bernoulli$(1/2)$ bit is drawn once per trajectory and reused at its action-token positions. If the bit is zero, all arms use q_0. For bit-one trajectories, the guidance is:

Arm	Intermediate decisions	Terminal decision
BASE	q_0	q_0
TERM-ONLY	q_0	q_1
NATURAL-FULL	Natural q_1	q_1
EARLY-PRIV	Early-answer adapter below	q_1

Terminal means the actual depth reached after five student continuations, including incorrect histories. If the student answers early, no terminal training position is added. NATURAL-FULL is the original formal experiment's FULL-PRIV condition with unchanged implementation; the renamed report label distinguishes it from the new intervention.

All fits use AdamW with learning rate 2×10^{-5}, weight decay 0, gradient clipping 1, batch 8, and a constant learning rate for 16 updates. The optimizer is initialized afresh for distillation. The loss sums full-vocabulary token-level forward KL over the student's sampled action tokens at temperature 1 and scales by $1/(8 \cdot 32)$; it is not normalized separately by trajectory length. The unprivileged calibration instead used supervised cross-entropy, so its learning curve is not a forecast of the distillation curve. Fitting seeds are $11, 22, 33, 44, 55$, checkpoints are $0, 1, 2, 4, 8, 16$, the primary endpoint is 8, and 16 is the secondary persistence endpoint. Teacher reference wording, early-answer mass, codebook, data, and endpoints were fixed before the new student runs.

For collection and evaluation, complete legal action strings are scored by summed full-vocabulary token log likelihoods and sampled after normalization across the legal candidates, at temperature 1 and without length normalization. The training student's softmax denominator always covers the full vocabulary. Free generation ends at its first answer or the six-action cap. Completion requires all five intermediate states to be correct before answering, independently of final-label correctness. Forced evaluation masks early answers but supplies no correct states and repairs no errors. Correct final answers after an incorrect or incomplete path still count in free accuracy.

Correctness given continuation is the number of correct emitted continuation actions divided by all emitted continuation actions. Per-depth values condition on actually reaching the depth and choosing to continue; absent observations remain missing. This sampled conditional metric is distinct from the supplementary codebook diagnostic on supplied correct terminal states. Invalid-action and truncation counts can be structural zeros under the action grammar.

A.1.3 THE EARLY-ANSWER TEACHER ADAPTER

At an intermediate decision let $C_1, \ldots, C_8$ be the complete legal continuation strings and A_Y the correct answer string. From the raw frozen q_0, compute each sequence likelihood $Q_0(C_i)$ using its token conditionals. The specified action distribution is

$$Q_{\text{early}}(C_i) = 0.8 \, \frac{Q_0(C_i)}{\sum_{k=1}^{8} Q_0(C_k)}, \qquad Q_{\text{early}}(A_Y) = 0.2. \tag{A.1}$$

All other complete action strings receive zero mass. This preserves the relative probabilities of *complete legal continuations*, not the raw token probabilities within them. No end-of-sequence token is appended, matching the action scorer.

A trie over the nine supported strings yields the token targets. For an action-token prefix z, let $m(z)$ be the sum of sequence weights extending z. When $m(z) > 0$,

$$q_{\text{early}}(v \mid z) = \frac{m(zv)}{m(z)}. \tag{A.2}$$

The target is a full-vocabulary vector with zeros outside the supported trie edges. If a student-emitted prefix has zero mixture mass, the remaining tokens of that action use raw q_0; the rule resets at the next environment decision. The first incompatible emitted token is still supervised by the preceding supported-prefix target. Thus the on-policy teacher is defined on supported prefixes and by an explicit fallback elsewhere, rather than by a single finite-support sequence law after arbitrary student deviations.

At adapted positions the KL is evaluated as

$$\sum_{v : q_v > 0} q_v (\log q_v - \log p_v),$$

using the student's ordinary unmasked full-vocabulary probabilities. Zero-safe arithmetic avoids $0 \log 0$ numerical errors. All original supervised positions and their weights are retained. At the terminal decision, EARLY-PRIV and TERM-ONLY use the identical natural q_1 target for any given prefix. Their realized terminal exposure can differ because their students reach that decision at different rates.

Legal-sequence conditioning also changes off-grammar support and internal token targets. Consequently, the intervention comparison measures the effect of this specified intermediate teacher adapter, not a scalar stopping-logit change with every other token target held fixed. Its construction and terminal-target equality are part of the experimental definition.

Implementation checks. The teacher-only audit used the first 32 frozen training tasks, giving 160 intermediate and 32 terminal prompts, without test access or student updates. At each intermediate prompt, complete-action continuation mass, conditional continuation probabilities, and products of trie conditionals matched the specified mixture within 10^{-6}. Terminal q_1 vectors agreed exactly. The natural q_0 and q_1 both assigned continuation probability almost one at intermediate depths; on the same diagnostic set the correct terminal-answer probability was 0.612835 for q_0 and 0.868031 for q_1. These probabilities are normalized complete-action probabilities, not raw first-token mass. Zero-support, finite-gradient, and unchanged-interface tests passed. The three unchanged seed-11 arms were rerun and checked against the archived formal runs at all six checkpoints before EARLY-PRIV fitting; all five seeds of all four arms were then reported. The three replay fits are included in the 20 fits, not counted twice.

A.1.4 Primary results and paired comparisons

Table 2 reports every primary-endpoint result. Means and sample standard deviations describe variation across five fits conditional on one shared calibration and task set. They do not represent five independently calibrated models. All signs in Table 3 are paired by fitting seed, using the same test questions.

Table 2: Every fitting seed at the prespecified update 8. Values are percentages. "Step" is correctness among emitted continuation actions, not a gold-prefix readout.

Arm	Seed	Forced	Completion	Free	Step	Early answer
BASE	11	57.81	100.00	58.98	100.00	0.00
BASE	22	59.38	100.00	57.03	100.00	0.00
BASE	33	57.42	100.00	60.16	100.00	0.00
BASE	44	53.52	100.00	51.17	100.00	0.00
BASE	55	52.34	100.00	53.91	100.00	0.00
TERM-ONLY	11	63.28	100.00	62.89	100.00	0.00
TERM-ONLY	22	72.66	100.00	71.88	100.00	0.00
TERM-ONLY	33	71.88	100.00	69.92	100.00	0.00
TERM-ONLY	44	69.53	100.00	70.31	100.00	0.00
TERM-ONLY	55	57.81	100.00	57.81	100.00	0.00
NATURAL-FULL	11	63.67	100.00	64.45	100.00	0.00
NATURAL-FULL	22	75.39	100.00	75.78	100.00	0.00
NATURAL-FULL	33	71.48	100.00	69.92	100.00	0.00
NATURAL-FULL	44	71.09	100.00	71.09	100.00	0.00
NATURAL-FULL	55	57.81	100.00	59.77	100.00	0.00
EARLY-PRIV	11	64.06	69.92	48.05	100.00	30.08
EARLY-PRIV	22	64.45	37.11	32.03	100.00	62.89
EARLY-PRIV	33	58.20	60.55	40.23	100.00	39.45
EARLY-PRIV	44	58.59	53.12	39.06	99.90	46.48
EARLY-PRIV	55	64.06	68.36	48.05	100.00	31.64

TERM-ONLY improves forced accuracy relative to BASE in all five seeds without a sampled completion loss. NATURAL-FULL has the same completion outcome. EARLY-PRIV lowers completion relative to TERM-ONLY in every seed, while its forced accuracy remains higher than the matched BASE in every seed. This yields the three-way signed contrast reported in the main text.

Table 3: Paired update-8 differences, in percentage points: mean ± sample standard deviation. The last column counts positive forced, negative completion, and negative free differences, respectively; a zero difference does not count toward a sign.

Comparison	Forced	Completion	Free	Signs
TERM − BASE	$+10.94 \pm 5.09$	$+0.00 \pm 0.00$	$+10.31 \pm 6.72$	5/5, 0/5, 0/5
NATURAL − TERM	$+0.86 \pm 1.28$	$+0.00 \pm 0.00$	$+1.64 \pm 1.47$	3/5, 0/5, 0/5
EARLY − TERM	-5.16 ± 8.38	-42.19 ± 13.38	-25.08 ± 12.42	2/5, 5/5, 5/5
EARLY − BASE	$+5.78 \pm 3.92$	-42.19 ± 13.38	-14.77 ± 7.62	5/5, 5/5, 5/5

The smallest positive forced contrast is 0.78 points (seed 33); the complete seed table conveys the variation rather than only the aggregate direction.

What accounts for execution failures? Across the five EARLY-PRIV primary test runs there are 1,280 free trajectories. Of these, 539 first fail to complete because they answer early, one because of an incorrect continuation, and 740 complete all five steps. Among the completed trajectories, 455 answer correctly and 285 give the wrong final label. A premature answer may itself be correct, so execution-failure counts are not the complement of free accuracy. The mean conditional continuation correctness is 99.9800%. This measurement conditions on the actions the student emits; depth-wise diagnostics are shown in Figure 5.

A.1.5 Training curves, depth-wise behavior, and fallback usage

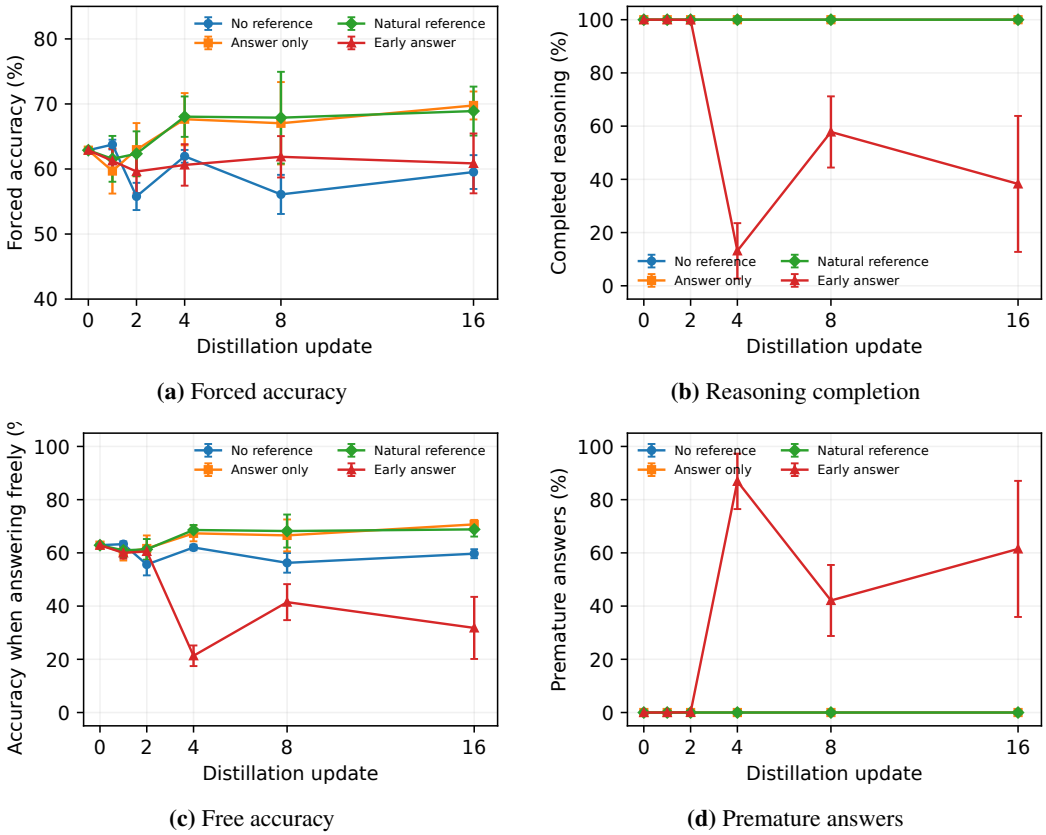
Figure 4 retains every scheduled checkpoint. The execution trajectory is non-monotone: EARLY-PRIV's mean completion is 13.12% at update 4, 57.81% at update 8, and 38.28% at update 16. These later or earlier points do not replace the primary comparison. At update 16, EARLY-PRIV still has lower completion and free accuracy than BASE in every seed, but its forced-accuracy contrast is positive in three of five seeds. The primary signed result is therefore an update-8 result, not a claim of a monotone or uniformly persistent accuracy separation.

The shared terminal teacher target does not imply the same number of terminal training positions or the same learned terminal predictor. EARLY-PRIV has 5.16 points lower forced accuracy than TERM-ONLY on average at update 8, with mixed paired signs. Reduced terminal exposure is one possible explanation consistent with the student-side analysis, but these runs do not separately identify it from shared-parameter and teacher-target effects.

Fallback use is recorded for all supervised tokens, not inferred from evaluation outcomes. Through update 8, the raw-q_0 fallback accounts for 0.37–1.06% of all supervised action-token positions in the five EARLY-PRIV fits. Its per-seed counts and denominators are shown below. At update 0 no such tokens have been supervised, so the corresponding rate is undefined rather than zero.

Seed	All supervised tokens	Adapter tokens	Fallback tokens	Fallback (%)
11	1838	884	18	0.979
22	1887	1121	20	1.060
33	1831	835	14	0.765
44	1775	568	15	0.845
55	1887	837	7	0.371

Reproducible evidence. The source package retains the supplied protocol, target implementation, audits, all 120 metric cells, paired comparisons, depth and fallback tables, and primary/persistence per-question records under `evidence/teacher_policy_20260916/`. The primary figures and tables are generated from these archived values. No additional model run, intervention-strength search, or endpoint selection was performed during manuscript integration.



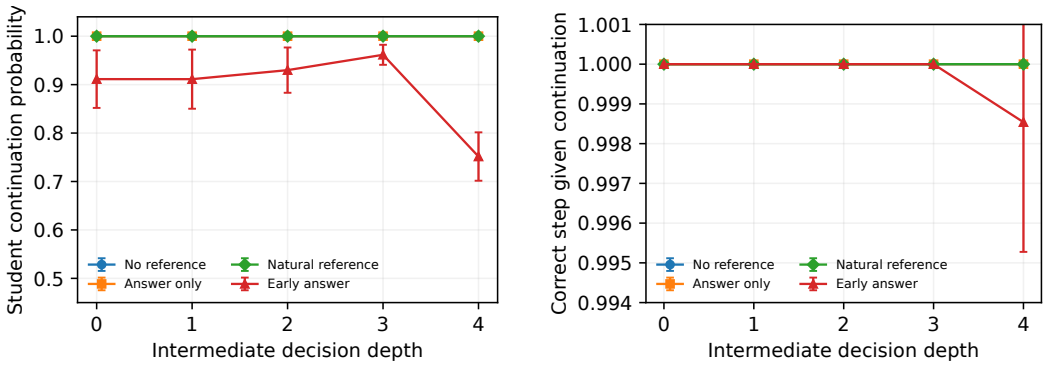
(a) Forced accuracy

(b) Reasoning completion

(c) Free accuracy

(d) Premature answers

Figure 4: Qwen3 training curves at all six scheduled checkpoints. Markers and error bars are means and sample standard deviations over five fits. Lines connect observations and do not assert behavior between checkpoints. The primary endpoint is update 8; update 16 is the persistence check.



(a) Whether the student continues

(b) Correctness when it continues

Figure 5: Depth-wise diagnostics at update 8, averaged over five fits with sample-standard-deviation bars. Continuation probabilities are measured on reached contexts; correctness additionally conditions on an emitted continuation. The scales differ because continuation changes substantially while observed step correctness stays near one.

A.2 Earlier experiments and extended results

A.2.1 Experimental setup

All language-model experiments use Qwen3-1.7B (Qwen Team, 2025) in BF16 and full-vocabulary forward KL at temperature 1. The earlier full-context composition experiment uses a dedicated shared checkpoint and constrained action grammar for trajectory sampling/evaluation; Study A and Study B use the fixed natural-language prompts used in the corresponding runs. Sampling truncation never truncates KL support.

Earlier full-context Qwen3 experiment. The controlled task draws fresh permutation-composition instances with $M = 8$ and continuation depth $D = 5$. Each output decision is either an explicit continuation action carrying the next state or a terminal answer action; a wrong continuation literally determines the state used by the next lookup. The shared checkpoint was calibrated before distillation on controlled per-action correct-prefix examples rather than concatenated full trajectories. We selected update 128 as the earliest checkpoint that passed the predefined behavior gate on a held-out calibration set: valid parsing, reliable no-reference completion and forced accuracy, high reference-conditioned answer accuracy, and earlier answering with the reference. On the 64-task calibration set, this checkpoint had no-reference completion 0.953, forced accuracy 0.984, reference accuracy 1.000, and first-answer depths 5 without the reference versus 0 with it. Later calibration checkpoints were retained but not used for the distillation experiments. Student and frozen teacher share the selected checkpoint; the student always receives no reference. A Bernoulli$(1/2)$ privilege bit is drawn once per student trajectory and reused at all prediction positions. The loss is token-level forward KL over the full Qwen vocabulary; grammar constraints are used only to sample/evaluate legal actions. The seven evaluated checkpoints are updates $0, 8, 16, 32, 64, 96, 128$ on one fitted trajectory. A fixed 256-task held-out set is used at every checkpoint. Forced evaluation masks early answer actions while retaining the student's own continuation choices; it never injects gold intermediate states.

The 256-task evaluation set was frozen after pilot training had begun; it was not used to change the model, data, or optimization settings. Checkpoint 0 is the exact pre-update model state evaluated retrospectively. The decision to extend training from update 64 to 128 was made after inspecting update-64 results. The controlled experiment is therefore a mechanistic illustration with task-conditional uncertainty, not a preregistered training-seed study. The experimental continuation depth $D = 5$ is an interface variable and is not identified with the paper's theoretical K notation.

Study A. Study A uses Qwen3-1.7B with LoRA rank 8/scaling 16/dropout 0 on attention q, k, v, o and MLP gate/up/down projections, and AdamW with learning rate 2×10^{-5}, weight decay 0, and gradient clipping 1. ON and PROCESS each run for 32 updates with effective batch 8 and training generation cap 2048. A reference bit with probability $1/2$ is drawn once per trajectory. ON uses the selected teacher throughout. PROCESS uses the unreferenced teacher through prediction of the literal `</think>` token and the selected teacher thereafter. Fitting seeds are $11, 22, 33, 44, 55$; the shared BASE checkpoint is used as a fixed reference. Each fitted model is evaluated on the same 96 GSM8K questions with one 8192-token maximal stream per question, whose prefixes define 2048/4096/8192 budgets. The earlier MATH-500 evaluation uses only the seed-11 fitted models and is treated as a single-seed external-task check.

Study B. For every training and development question, one frozen-teacher trajectory is collected without a reference and one with a reference: 256 training plus 96 development trajectories. RETAIN reuses the collection-time reference indicator. RELABEL independently samples a supervision bit once per trajectory and reuses it throughout training. Both arms see the same cached prefixes and use 64 updates at batch 4 with fitting seeds 11 and 22 on one shared collection cache. The loss samples stratified positions with inverse-inclusion weights and full prefix context. Seeds change fitting/relabeling randomness but do not recollect trajectories.

Uncertainty. Controlled-Qwen task intervals condition on one fitted training trajectory and the fixed 256 tasks. Study-A question-bootstrap intervals condition on each fitted model; the five fitted seeds are shown individually, and their means, sample standard deviations, and medians are descrip-

tive rather than high-precision training-seed population estimates. Study-B intervals condition on the shared collection cache. Intervals spanning zero indicate unresolved differences, not equivalence.

A.2.2 Earlier full-context Qwen3 results

This experiment previously appeared in the main text; it is retained here with its original results. Unlike the staged-reveal experiment, every prompt exposes all lookup tables, so completion measures the emitted reasoning path. Table 4 reports its metrics at all seven real checkpoints. The sharp shift occurs early: free accuracy and completion collapse after eight updates, while forced accuracy remains substantially higher throughout training.

Table 4: Controlled Qwen3 primary metrics on the fixed 256-task set.

Update	Free accuracy	Forced accuracy	Completion
0	0.9688	0.9844	0.9688
8	0.1094	0.9219	0.0039
16	0.1328	0.8359	0.0078
32	0.1641	0.6523	0.0625
64	0.1484	0.7266	0.0273
96	0.1172	0.6953	0.0039
128	0.1641	0.7344	0.0391

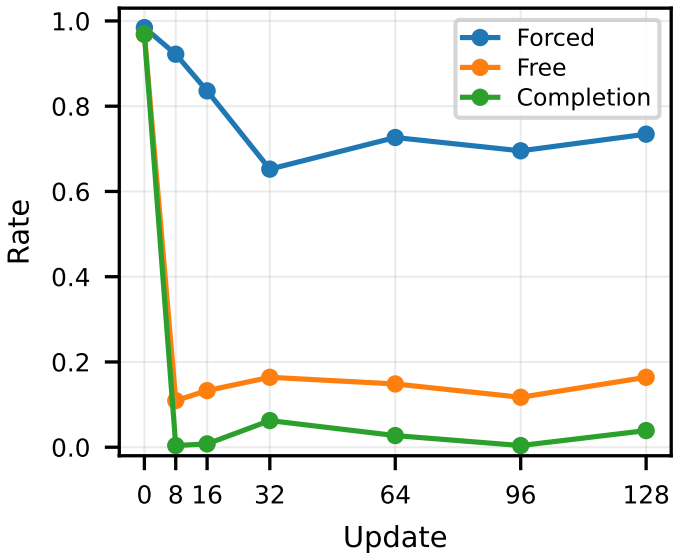


Figure 6: The earlier full-context Qwen3 experiment, retained from the previous main-text figure. Free accuracy and explicit completion fall sharply; forced accuracy also declines from its near-ceiling initialization. These are the seven original measured checkpoints.

Failure mode and conditional computation. At update 128, the first failure event is a premature answer on 242 trajectories and a wrong continuation on 4, with 10 trajectories reaching the terminal decision without error and no malformed actions. Thus 98.37% of execution failures first manifest as premature answering. On correct prefixes, the conditional probability of the correct continuation given a continuation choice remains high (approximately 0.95–0.99 across the five depths at update 128), whereas the mass assigned to continuing is much lower. The controlled task therefore separates conditional lookup competence from the policy to continue executing it.

Incomplete-trajectory baseline check. Because $M = 8$, the unresolved baseline is $u = 1/8 = 0.125$. After distillation, answer accuracy on incomplete trajectories stays close to this baseline:

0.1260 at update 16, 0.1250 at 32, 0.1245 at 64, and 0.1301 at 128. The empirical partition satisfies

$$A_{\text{free}} = P_{\text{complete}} A_{\text{complete}} + (1 - P_{\text{complete}}) A_{\text{incomplete}}.$$

For

$$A_{\text{proxy}} = u + (A_{\text{complete}} - u) P_{\text{complete}}, \qquad A_{\text{free}} - A_{\text{proxy}} = (1 - P_{\text{complete}})(A_{\text{incomplete}} - u).$$

Figure 7 therefore primarily checks whether incomplete trajectories behave like unresolved guesses; it is not an independent validation of Theorem 1.

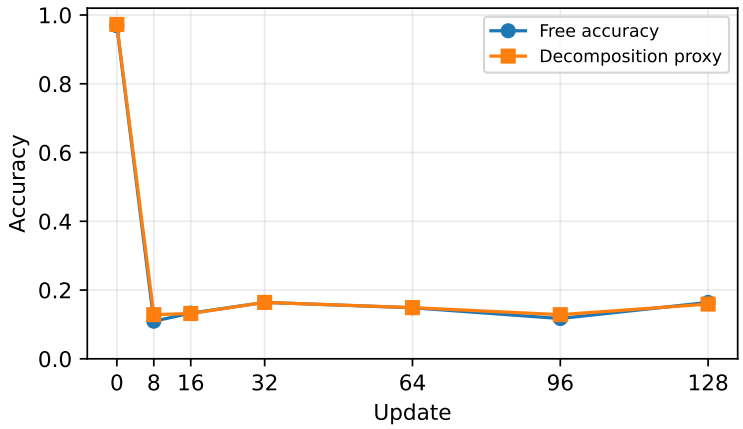


Figure 7: Observed free accuracy and the baseline-substitution proxy across the seven controlled-Qwen checkpoints. Their closeness reflects that incomplete-trajectory accuracy stays near $u = 1/M$; it is a descriptive check rather than an independent theorem prediction. Connecting lines are visual guides only.

Terminal readout and exposure. On the correct terminal prefix, the mean probability of the correct answer is 0.99999 at update 0, 0.7223 at update 32, 0.7479 at 64, 0.7953 at 96, and 0.7785 at 128. Thus readout quality worsens sharply and then partially, non-monotonically recovers; it never exceeds the near-perfect initial checkpoint. Figure 8 retains the complete seven-point diagnostic. Terminal supervised action positions also become sparse: 43 occur in updates 1–8, compared with 28 over updates 9–64 and 22 over updates 65–128. We use these counts only as an empirical visit/exposure analogue; the measured terminal error is not monotonic in this exposure summary.

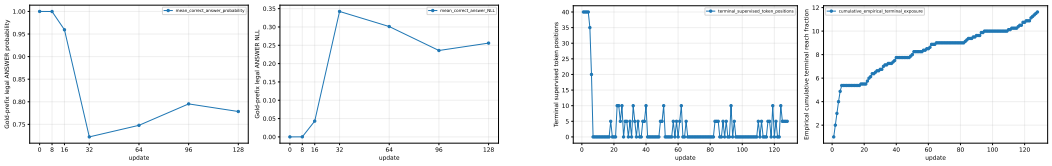


Figure 8: Controlled-Qwen auxiliary diagnostics. Left: correct-prefix terminal readout degrades and partially recovers non-monotonically. Right: terminal/deep training exposure becomes sparse after the execution shift.

A.2.3 Numerical results on the composition task

The synthetic evaluation includes 45 finite-feature checks and analytic/Monte Carlo comparisons. To examine the finite-training separation in Theorem 2, we use the S2 configuration with $K = M = 2$, $c_0 = 0.99$, $c_{\text{ref}} = 0.10$, $a_0 = 0.90$, $a_{\text{ref}} = 0.99$, $\rho = 0.20$, $\eta = 1$, $N = 600$, and $L = 20$. Free accuracy under the finite-population dynamics decreases from 0.896000 to 0.839416 while forced accuracy increases from 0.900000 to 0.918000. Continuation falls to 0.812000 and terminal accuracy rises to 0.918000.

The run illustrates the separation at a moderate suppression parameter. The theorem's conservative sufficient condition instead gives $N_{\text{sufficient}} = 266$ and $L_{\text{sufficient}} \simeq 1484.812$, whereas the run uses $L = 20$ and therefore does not satisfy this sufficient bound. Figure 9's separate comparison-clock run uses $M = 8, K = 6, c_0 = 0.97, c_{\text{ref}} = 0.35, a_0 = 0.8, a_{\text{ref}} = 0.95, \rho = 0.3, \eta = 1$, and 600 updates. Its terminal target is $a_* = 0.845$. The plotted error is $a_* - a(N)$ against the recorded H_N; this clock belongs to the comparison dynamics, rather than the exact finite network.

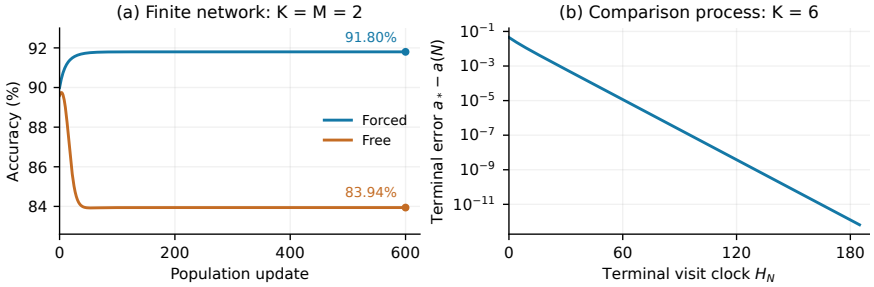


Figure 9: Completed controlled runs. Left: the actual finite-softmax free/forced separation. Right: terminal error versus its visit clock in a separate comparison process. Each panel uses 600 recorded updates; the right panel is not a finite-network measurement.

To examine the protocol formulas in Theorem 3, the protocol simulation uses $M = 8$, $K = 5$, $\rho = 0.4$, $(c_0, c_1) = (0.95, 0.45)$, $(a_0, a_1) = (0.8, 0.95)$, seed 2026, and 30,000 collection plus 30,000 fresh evaluation samples. RETAIN and RELABEL accuracies are 0.467400 and 0.356033, against analytic values 0.468407 and 0.357559. These are finite-policy checks, not neural training from scratch.

A.2.4 Study A across five fitted seeds

Table 5 gives the five fitted-seed GSM8K results. The shared BASE scores 59.38/78.12/84.38% at 2048/4096/8192 tokens and uses 2314.08 mean thinking tokens at 8192.

Table 5: Study A GSM8K results for the five fitted seeds. Each row uses one 8192-token stream per question; 2048/4096 scores are prefixes of that same stream.

Seed	Arm	2048	4096	8192	Mean thinking@8192
11	ON	75.00	82.29	86.46	1669.63
11	PROCESS	62.50	79.17	89.58	2147.49
22	ON	70.83	80.21	83.33	1512.09
22	PROCESS	58.33	77.08	85.42	2211.91
33	ON	75.00	83.33	85.42	1514.30
33	PROCESS	54.17	78.12	83.33	2235.92
44	ON	72.92	79.17	83.33	1679.51
44	PROCESS	56.25	77.08	88.54	2313.14
55	ON	73.96	81.25	84.38	1614.66
55	PROCESS	61.46	78.12	89.58	2246.84

Across fitted seeds, ON–BASE thinking-length effects are $-644.46, -801.99, -799.78, -634.57, -699.43$ tokens (mean -716.05, sample SD 81.30); all five are negative. PROCESS–ON effects are $+477.86, +699.81, +721.61, +633.63, +632.19$ tokens (mean $+633.02$, sample SD 95.37); all five are positive. Figure 10 retains the individual same-stream budget curves. Long-budget ON accuracy varies from 83.33% to 86.46%, centered near BASE at 84.38%; the robust effect is therefore on execution length rather than long-budget harm.

A question-level comparison helps interpret that shortening. In every fitted seed, roughly four fifths of the 96 questions are correct under both BASE and ON, and ON is shorter on most of these correctness-preserving cases. Correct-to-wrong transitions are much less common. Table 6 summa-

rizes the BASE-to-ON comparison. Because the models are stochastic fitted policies and length is not experimentally manipulated, this is an association rather than a causal test of stopping.

Table 6: Question-level BASE-to-ON transitions on GSM8K. "Shorter" and mean shortening are computed within the questions that remain correct under both models. The final column gives the number of correct-to-wrong transitions out of 96 questions.

Seed	Correct→correct	Shorter (%)	Mean shortening	Correct→wrong
11	79	92.4	546	2
22	77	90.9	595	4
33	78	92.3	657	3
44	76	86.8	518	5
55	77	90.9	605	4

Across seeds, the correct-to-wrong cases do not show consistently greater shortening than the correct-to-correct cases. The reverse ON-to-PROCESS comparison is similarly heterogeneous: PROCESS is usually longer on questions that remain correct, but extra length does not guarantee an accuracy improvement. These paired comparisons are therefore consistent with substantial compression on many questions while leaving open a smaller subset in which length and correctness change together.

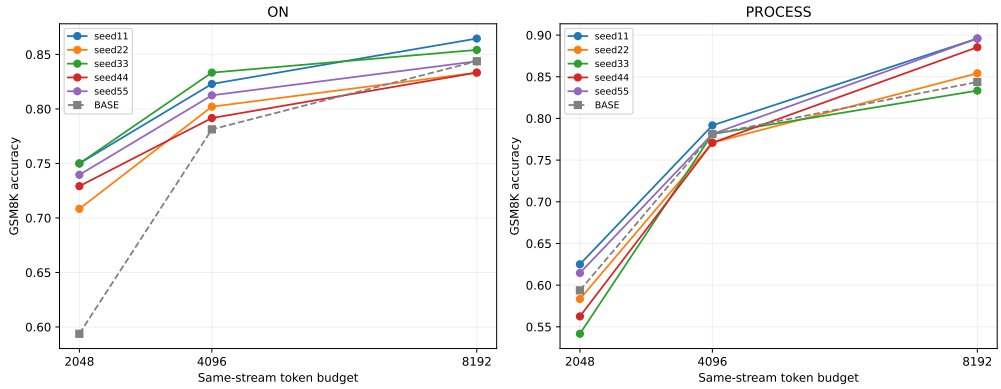


Figure 10: Study A GSM8K accuracy versus same-stream generation budget for all five fitted seeds. Individual seeds remain visible; BASE is the shared BASE reference.

ON's observed 2048-to-8192 gain is 10.42–12.50 points across seeds, compared with 25.00 points for BASE. PROCESS gains 27.08–32.29 points and exceeds ON's observed gain in all five seeds. These increments depend on starting accuracy, remaining headroom, and truncation and cannot by themselves measure useful computational ability.

Single-seed external-task check. The earlier fixed 64-question MATH-500 result is reported only for seed 11: BASE/ON/PROCESS score 23.44/42.19/20.31% at 2048 tokens, 73.44/78.12/68.75% at 4096, and 81.25/90.62/87.50% at 8192. ON's positive 8192-token result is a measured single-seed benefit, not five-seed robustness evidence.

Local coordinates and conditional readout. The seed-11 closing probe selects 40 naturally closed trajectories from a 64-question BASE development pool, yielding 199 shared prefixes. ON–BASE mean log closing probability is $+2.991$ ($[2.697, 3.285]$), but mean raw probabilities are 1.07×10^{-16} for BASE and 2.73×10^{-15} for ON. The close-plus-EOS contrast spans zero. The absolute scale therefore leaves a functional stopping effect unestablished. On fixed BASE prefixes, ON reduces conditional gold-answer NLL relative to BASE on both the eight natural-closure and eight artificial-closure groups. This fixed-prefix quantity is a readout proxy, not forced rollout accuracy.

Deferring termination. On the fixed 32-question subset in Appendix A.3.2, the accuracy difference-in-differences is -3.125 points ($[-9.375, 0]$) at all three budgets, with 31/32 BASE and

30/32 ON streams unchanged. The intervention does not identify an ON-specific rescue and is retained as a negative mechanism check.

A.2.5 STUDY B: COLLECTION SELECTION AND RELABELING

Figure 11 shows the teacher-collection effect. At depth 1024, the unreferenced and referenced risk sets contain 86/128 and 27/128 training trajectories; the survival difference is -0.46094 ($[-0.55469, -0.36719]$), and the referenced fraction among survivors is $27/113 = 0.23894$ ($[0.17757, 0.29630]$), below the balanced initial fraction $1/2$. Development counts show the same direction (31/48 versus 6/48).

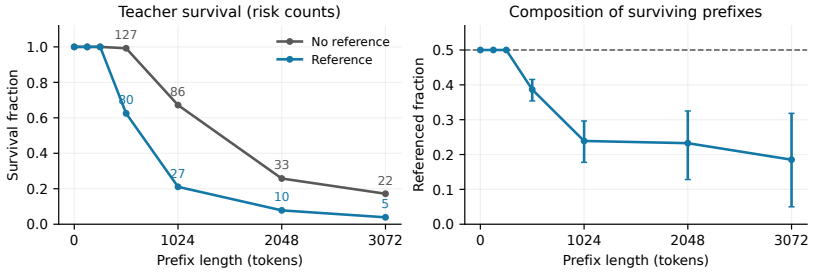


Figure 11: Study B: referenced frozen-teacher trajectories survive less often to deeper prefixes, reducing their share among surviving trajectories. Numbers mark the paired training risk sets.

A trajectory remains at risk if it has a predictable position after the specified number of tokens and no prior close/EOS. The cap is administrative censoring, and the empirical curves need not have the constant hazards of Theorem 3. The reduced referenced fraction establishes selection in the collected prefixes; consequences for the fitted student are analyzed through the target and local-policy diagnostics below.

Supervision targets. For a collected prefix h, let $g_b(h) = q_b(\texttt{</think>} \mid h)$ and $d(h) = g_1(h) - g_0(h)$. On inverse-inclusion-weighted selected pre-close positions with $512 \le t < 2048$,

$$\begin{aligned}\mathbb{E}[g_{\mathrm{REL,mean}} - g_{\mathrm{RET}}] &= \mathbb{E}[(\tfrac{1}{2} - B_{\mathrm{col}})d] \\ &= (\tfrac{1}{2} - \mathbb{E}B_{\mathrm{col}})\mathbb{E}d - \mathrm{Cov}(B_{\mathrm{col}}, d). \end{aligned} \tag{A.3}$$

Here $\mathbb{E}B_{\mathrm{col}} = 0.26240$ and the resulting expected-mixture gap is -3.64×10^{-10} ($[-5.06, -2.39] \times 10^{-10}$). The sign is not determined by the selected referenced fraction alone, and the very small absolute scale leaves a functional stopping effect unestablished.

Student local coordinates and free generation. On shared development prefixes, RELABEL has a lower mean log closing probability than RETAIN in both fitted seeds, but raw probability differences remain on the order of 10^{-15} and close-plus-EOS intervals span zero. At 4096 tokens, paired RELABEL–RETAIN accuracy differences are $+3.13$ and $+1.04$ points for seeds 11 and 22, with both intervals spanning zero; thinking-length differences also change sign. Thus the collection-selection effect is clear, whereas a stable macro-level protocol ordering remains unidentified. Full local-coordinate tables are in Appendix A.3.3.

A.2.6 LIMITATIONS

The controlled Qwen3 run follows one fitted training trajectory and uses a held-out set frozen after the pilot had begun; training was extended from 64 to 128 updates after inspecting the update-64 results. Study A now covers five fixed fitted seeds, but each fitted model still contributes one sampled evaluation stream per question and the shared BASE is not five independent fits. Study B uses two fits of one shared collection cache rather than independently recollected caches. Thinking length is a literal token count, not a validated measure of useful computation. These limitations motivate the paper's interpretation as controlled mechanistic evidence plus seed-aware behavioral replication, not a universal causal claim about privileged distillation.

A.3 Additional empirical details

A.3.1 Assets, exposure, and estimands

All language-model experiments use Qwen3-1.7B. Study A evaluates GSM8K and MATH-500, while the lookup studies use separate permutation-composition generators. The staged-reveal teacher-policy study is documented in Appendix A.1; references below to the earlier lookup run concern the full-context experiment. The natural-language studies use distinct training pipelines and are not intended as matched training-FLOP comparisons.

For the Study-A seed-11 run, ON and PROCESS expose 30,607 and 28,157 post-thinking privileged tokens, with cumulative update-norm summaries 1.0920 and 0.6859. These are descriptive implementation diagnostics, not separately randomized causes; the newer fitted seeds use the same scientific protocol but have their own on-policy rollouts. Study B freezes the base embeddings and language-model head and uses a fixed cache-average length in the KL denominator. In neither natural-language study is the training objective a per-trajectory normalized average.

Controlled-Qwen completion is an explicit action-level event; GSM8K/MATH accuracy instead measures correct final boxed answers under the fixed parser and thinking-boundary rule. Natural-language thinking length is a literal token count, not a measure of useful computation. Conditional answer NLL and gold-prefix readout test likelihood on supplied prefixes. The close-coordinate probes measure probabilities at selected fixed prefixes and do not estimate whole-rollout stopping probability. Budget increments compare truncations of the same sampled stream. These estimands are kept separate throughout the paper.

Statistical scope. The earlier full-context Qwen intervals condition on one fitted trajectory. Its evaluation set was frozen after pilot training began; model/data/optimization settings were unchanged, but the 64-to-128 extension followed inspection of update 64 and is exploratory. Study-A question-bootstrap intervals condition on each fitted model; all five seeds are shown individually. Study-B fitted-seed comparisons condition on one shared teacher-collection cache.

A.3.2 Deferred-termination four-cell comparison

Condition (same 32 questions)	2048	4096	8192
BASE free	75.000	84.375	90.625
ON free	87.500	90.625	90.625
BASE defer	71.875	81.250	87.500
ON defer	81.250	84.375	84.375
Difference-in-differences (points)	-3.125	-3.125	-3.125

Each difference-in-differences interval is $[-9.375, 0]$ points. Free cells here come from the intervention's 32-question subset, not the 96-question headline evaluation. The original cell intervals, correct/incorrect transitions, and unchanged-stream checks remain in the accompanying evidence tables and archives.

A.3.3 Study B evaluation and local coordinates

	Accuracy (%)		Mean thinking tokens	
Model	2048	4096	2048	4096
BASE	62.50	82.29	1409.91	1800.15
RETAIN SEED11	71.88	83.33	1163.03	1396.75
RELABEL SEED11	75.00	86.46	1159.76	1414.98
RETAIN SEED22	72.92	85.42	1183.02	1419.36
RELABEL SEED22	73.96	86.46	1122.91	1352.82

Seed	RELABEL–RETAIN, 4096 tokens	Estimate	95% CI
11	Accuracy (points)	3.12	$[-1.04, 7.29]$
11	Thinking tokens	18.23	$[-80.57, 113.40]$
22	Accuracy (points)	1.04	$[-3.12, 5.21]$
22	Thinking tokens	-66.54	$[-172.49, 43.13]$

The fitted-seed contrast intervals condition on the same collection cache. The two seeds are not treated as two independent collection experiments. No pooled seed confidence interval is constructed from their arithmetic average.

A.3.4 OTHER DIAGNOSTICS

The initial common-prefix teacher probe and one-step checks are auxiliary diagnostics. They are not interchangeable with the later trained-student closing probe. The teacher's access to the full solution is privileged and does not define a fair benchmark competitor. Dense budget grids reuse existing streams, and heuristic reasoning markers do not identify semantic harm. Format failures, cap hits, conditional NLL, and teacher-quality diagnostics are retained in the machine-readable evidence tables.

B THEORETICAL SETTINGS AND ASSUMPTIONS

The main text gives only the assumptions needed to read the results. Here we collect the full theorem settings, notation, mechanism-level decompositions, and finite-horizon constants used in the proofs.

B.1 TASK, EVALUATION, AND FITTING DEFINITIONS

We give the precise notation for the behavioral description in Section 3. The three settings below share the task and teacher-reference interface but differ in their readout classes and fitting or update rules.

Composition task. Each task asks the model to predict the final state of a sequence of input mappings. Let $[M] = \{1, \dots, M\}$, let $\mathfrak{S}_M$ be its permutation group, and take integers $K, M \geq 2$. Each example independently draws

$$\begin{aligned} f_1, \dots, f_K \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathfrak{S}_M), \qquad S_0 \sim \mathrm{Unif}([M]), \quad S_0 \perp (f_1, \dots, f_K), \\ X = (f_1, \dots, f_K, S_0), \quad S_j = f_j(S_{j-1}), \quad Y = S_K. \end{aligned} \tag{B.1}$$

All mapping tables and S_0 are included in X, so $H(Y \mid X) = 0$. Write $d = K - 1$ and $u = 1/M$.

Generated states. Continuation tokens supply the state for the next lookup, whereas answer tokens terminate generation. The output vocabulary is

$$\mathcal{V} = \mathcal{C} \cup \mathcal{A}, \qquad \mathcal{C} = \{C_{j,s} : j \in [d],\ s \in [M]\}, \qquad \mathcal{A} = \{A_i : i \in [M]\}. \tag{B.2}$$

Starting from $(t, s) = (0, S_0)$, an emitted $C_{j,s}$ sets the literal stage and state to (j, s); the next lookup is $z = f_{t+1}(s)$. This rule also applies to incorrect or out-of-order continuations. On the prescribed full path, $t = d$ is the terminal stage and $z = Y$.

Restricted readouts. The constructions restrict which input information reaches the trainable readout, rather than removing information from the task. Frozen attention and ReLU maps give the unreferenced student features

$$F_1 = F_2 = (t, s, z, \perp), \qquad F_3 = (t, z), \qquad U_{\mathrm{raw}} = 0. \tag{B.3}$$

Here $U_{\mathrm{raw}} = 0$ excludes direct readout from raw residual and head coordinates. The restriction makes "having completed the computation" a controlled property of the readout: the full prompt still contains all tables, but the trainable output layer cannot bypass the intended intermediate computation through raw coordinates. Only the specified readout parameters are fitted or updated; the proof appendices give explicit realizations for all three theorem settings. In Setting 1, the supported continuation at an unfinished state is the unique correct next-state token, so Theorem 1 isolates the decision to execute the computation rather than errors within a chosen continuation.

Privileged teacher. Teacher and student share an initial checkpoint, but only the teacher receives an answer reference. For a generated prefix h, the frozen checkpoint p_0 supplies

$$B \sim \text{Bernoulli}(\rho), \qquad r_0 = \perp, \quad r_1 = Y, \qquad q_b(\cdot \mid X, h) = p_0(\cdot \mid X, h, r_b), \tag{B.4}$$

where $0 \le \rho < 1$ and $\perp$ denotes an empty reference. Each bit is drawn once per trajectory independently of the task; the student always uses $\perp$. Section 4.3 specifies the dependence between collection and labeling bits.

Evaluation. Free generation evaluates the complete sampled trajectory, while forced evaluation suppresses early answers without supplying correct intermediate states. For generated tokens V_m, free generation stops at its first answer or after K outputs:

$$T_A = \inf\{m \ge 1 : V_m \in \mathcal{A}\}, \qquad A_{\text{free}}(p) = \Pr_p(T_A \le K,\ V_{T_A} = A_Y). \tag{B.5}$$

A correct answer after an incorrect continuation counts as success; no answer by the cap is failure. For the finite-training setting of Theorem 2, forced evaluation uses

$$p_m^{\text{for}}(v \mid h) = \begin{cases} p(v \mid h)\mathbf{1}\{v \in \mathcal{C}\}/p(\mathcal{C} \mid h), & m \le d, \\ p(v \mid h), & m = K. \end{cases} \tag{B.6}$$

Here m is the emitted position, not the literal stage t. Theorem 1 instead masks answers while $t < d$. Write A_{forced} for forced accuracy and F_L, G_L for the finite-softmax free and forced accuracies in Theorem 2.

Distillation objective. Distillation matches full teacher distributions at the prediction positions visited by the collection policy. Let μ be the joint task, collection, and labeling law, and let $\mathcal{H}(\tau)$ contain the visited prediction prefixes, including the position at which the first answer is emitted, if any. Then

$$\mathcal{L}_\mu(p) = \mathbb{E}_\mu\left[\sum_{h \in \mathcal{H}(\tau)} D_{\text{KL}}\big(q^{\text{lab}}(\cdot \mid X, h) \,\big\|\, p(\cdot \mid F(X, h))\big)\right]. \tag{B.7}$$

Here B_{lab} denotes the privilege bit used for labeling, with the protocol-specific rule specified in Section 4.3. The loss sums over tokens without dividing by trajectory length.

The gate/content decomposition is given in (3.2).

Fitting and finite updates. The theorems distinguish population fits from finite training trajectories. For Theorems 1 and 3, Π is the appropriate independent-column probability-closure class in Settings 1 and 3, permitting exact zero probabilities; the first setting has prescribed structured supports, while the third allows full-simplex columns. A fitted policy satisfies

$$\pi \in \arg\min_{p \in \Pi} \mathcal{L}_{\mu[\pi]}(p). \tag{B.8}$$

For student collection, $\mu[\pi]$ is induced by π and held fixed during minimization; for teacher collection, it is independent of π. Theorem 2 instead uses the finite-softmax family $p_{\theta,L}$ of Setting 2, with suppression parameter $L > 0$ and strictly positive output probabilities, and updates only the tied coordinates $\theta = (z_1, \ldots, z_d, w)$:

$$\theta_{n+1} = \theta_n - \eta\, \nabla_\theta \mathcal{L}_{\mu[\theta_n]}(p_{\theta,L})\big|_{\theta=\theta_n}. \tag{B.9}$$

The expectation is exact, not a sampled batch; teacher targets, sampling law, and prefix contents are held fixed during differentiation.

B.2 Compact theorem-level settings

For a channel $Q_b \ge 0$ with $\sum_i Q_b(i \mid y) = 1$ for every y, abbreviate

$$\mathsf{Q}_b(y) = \sum_{i=1}^{M} Q_b(i \mid y)\delta_{A_i}, \qquad a_b = M^{-1}\text{tr}(Q_b), \qquad U_{\mathcal{A}} = M^{-1}\sum_{i=1}^{M}\delta_{A_i}. \tag{B.10}$$

Here δ_v is point mass at v, and $\text{logit}(x) = \log[x/(1-x)]$.

Setting 1 (Theorem 1: structured probability-closure fitting)**.** Use the finite-softmax attention and paired-ReLU construction specified in Appendix B.3. Empty-reference columns vary independently in the probability closure, supported on $\{C_{t+1,z}\} \cup \mathcal{A}$ for $t < d$ and in $\mathcal{A}$ for $t = d$. Reference columns are frozen, with checkpoint

$$q_0 = \begin{cases} \delta_{C_{t+1,z}}, & t < d, \\ \mathsf{Q}_0(z), & t = d, \end{cases} \qquad q_1 = \delta_{A_Y}, \qquad a_0 > u. \tag{B.11}$$

Use student-collected fitting (B.8) with $0 \le \rho < 1$; Appendix B.3 lists the full conditions.

Setting 2 (Theorem 2: finite softmax and tied population updates)**.** Use the finite-softmax construction specified in Appendix B.4, with $L > 0$ and trainable $\theta = (z_1, \dots, z_d, w)$. Its empty-reference logits are

$$\ell_\theta(v) = \begin{cases} z_{t+1} + \log M, & t < d,\ v = C_{t+1,z}, \\ w + \log(M-1), & t = d,\ v = A_z, \\ 0, & v \in \mathcal{A} \text{ in all remaining cases}, \\ -L, & \text{otherwise.} \end{cases} \qquad p_{\theta,L} = \operatorname{softmax}(\ell_\theta). \tag{B.12}$$

Reference coefficients and attention scale follow the same construction. Require

$$0 < c_{\mathrm{ref}} < c_0 < 1, \qquad u < a_0 < a_{\mathrm{ref}} < 1, \qquad 0 \le \rho < 1, \quad 0 < \eta \le 1, \tag{B.13}$$
$$\theta_0 = (\operatorname{logit}(c_0), \dots, \operatorname{logit}(c_0), \operatorname{logit}(a_0)).$$

These probabilities are nominal, after deleting suppressed rows. Update the K tied coordinates by (B.9); Appendix B.4 gives the complete assumptions.

Setting 3 (Theorem 3: hard attention and protocol-specific fits)**.** Use the hard-attention construction and independent probability-closure columns specified in Appendix B.6, with frozen reference columns and

$$\begin{aligned} &t < d: \quad q_0 = c_0 \delta_{C_{t+1,z}} + (1 - c_0) U_{\mathcal{A}}, \qquad q_1 = c_1 \delta_{C_{t+1,z}} + (1 - c_1) \mathsf{Q}_1(Y), \\ &t = d: \quad q_0 = \mathsf{Q}_0(z), \qquad q_1 = \mathsf{Q}_1(Y). \end{aligned} \tag{B.14}$$

Here $0 \le c_1 < c_0 \le 1$, $u < a_0 \le a_1 \le 1$, $0 \le \rho < 1$, and Q_0, Q_1 may be arbitrary column-stochastic matrices. Fit each protocol by (B.8); Appendix B.6 gives the complete protocol assumptions.

Two mechanism-level decompositions. The exact construction in Theorem 1 is a special case of a general stopped-trajectory identity. Let $P(\rho)$ be the probability of completing the designated computation, $a(\rho) = \Pr(\text{correct} \mid \text{complete})$, and $b(\rho) = \Pr(\text{correct} \mid \text{incomplete})$. Where these functions are differentiable,

$$A(\rho) = P(\rho)a(\rho) + [1 - P(\rho)]b(\rho), \qquad A'(\rho) = Pa' + (1 - P)b' + (a - b)P'. \tag{B.15}$$

Thus privilege may improve either completed-path prediction or shorter-path prediction while still reducing completion. Theorem 1 isolates the controlled specialization $b(\rho) = u$ and $a(\rho) = a_\rho$. In general, $a(\rho)$ need not equal forced accuracy; that correspondence uses the additional structure of the controlled setting.

The gate/content chain rule is stated in the main text in (3.2); it applies with the usual extended-KL convention at boundary masses.

The probability-closure settings of Theorems 1 and 3 permit exact zeros; every output probability in Theorem 2 is positive. Population fits and finite updates are analyzed separately.

B.3 Theorem 1 setting

We study a controlled neural construction in which privileged labels improve terminal answers while changing the probability that a student executes the required computation. The result concerns an explicit finite decoder, a restricted readout, and a conditional population fitting interface. Its execution comparison separates completion, forced continuation, and the first sampled answer. **Task and observation.** Write $[M] = \{1, \dots, M\}$, let $\log$ denote the natural logarithm, and let K, M be

integers. An example is $X = (f_1, \ldots, f_K, S_0)$, with $S_t = f_t(S_{t-1})$ for $1 \le t \le K$ and $Y = S_K$. The prompt contains every table record $(\ell, i, f_\ell(i))$, a reference slot $r \in \mathcal{R} = \{\perp\} \cup [M]$, and an initial query carrying $(0, S_0)$. The reference precedes the initial query. The output vocabulary is

$$\mathcal{V} = \{C_{\ell,i} : 1 \le \ell \le K-1,\ i \in [M]\} \cup \{A_i : i \in [M]\}, \qquad |\mathcal{V}| = KM.$$

A continuation $C_{\ell,i}$ literally changes the current stage/state to $(t, s) = (\ell, i)$; this rule applies even after an incorrect or out-of-order continuation. The next retrieved state is then $u = f_{t+1}(s)$. A correct prefix of length t is $C_{1,S_1}, \ldots, C_{t,S_t}$, including the empty prefix at $t = 0$. Full X determines Y; restricted readout information, rather than partial observation of X, is the object of study.

Architecture specification. Use one causal attention layer, exactly two heads, one residual MLP with a single hidden ReLU layer, no LayerNorm, and one shared output matrix U. Table-key, table-value, current-query, reference-value, constant, and decoded-feature coordinates occupy disjoint blocks. Embeddings encode only their primitive table record or token. The current query has stage/state one-hots $h^{\text{stage}} = e_t \in \mathbb{R}^K$, $h^{\text{state}} = e_s \in \mathbb{R}^M$. The step-head query is (e_{t+1}, e_s); table record $(\ell, i, f_\ell(i))$ has key (e_ℓ, e_i) and value $e_{f_\ell(i)} \in \mathbb{R}^M$. Non-table step keys and values are zero. The reference head uses a constant query, key 1 only at the reference slot and key 0 elsewhere, with value $e_r \in \mathbb{R}^{M+1}$ only at that slot. Both heads use softmax of these scores multiplied by a finite scale λ; ordinary attention normalization factors, if present, are absorbed into the query/key weights. Let $v^{\text{step}}, v^{\text{ref}}$ be their returned value vectors and set $N_{\max} = KM + K + 1$.

For each tuple $z = (t', s', u', r')$ in

$$\mathcal{Z} = \{0, \ldots, K-1\} \times [M] \times [M] \times \mathcal{R},$$

define the MLP coordinate

$$b_z = h^{\text{stage}}_{t'} + h^{\text{state}}_{s'} + v^{\text{step}}_{u'} + v^{\text{ref}}_{r'}, \qquad \phi_z = 2\big[\mathrm{ReLU}(b_z - 13/4) - \mathrm{ReLU}(b_z - 15/4)\big].$$

The two hidden units per tuple write into initially zero, otherwise unused residual coordinates. There are $2KM^2(M+1)$ hidden units and $KM^2(M+1)$ decoded coordinates, plus $O(K+M)$ primitive embedding/head coordinates. Retrieval and MLP weights are frozen. U reads only the ϕ block: all its coefficients on raw residual and head coordinates are fixed at zero. The proof below shows that this construction gives exactly $\phi = e_{(t,s,f_{t+1}(s),r)}$; the readout ignores the unused raw attention coordinates.

Closure policy class and common checkpoint. Once exact tuple decoding is established, identify each column of U with its output distribution indexed by z. Let $\mathcal{P}$ be the probability closure of such column readouts with these supports: for empty-reference $t < K-1$, only $C_{t+1,u}$ and all A_j may have positive mass; at empty-reference $t = K-1$, only answers may have positive mass. Nonempty-reference columns are frozen to the checkpoint below. The row $C_{t+1,u}$ is determined by the decoded tuple. Closure probabilities are interpreted as limits of finite softmax columns, so zero mass need not correspond to a finite logit. Define a column-stochastic matrix $Q_0 \in [0,1]^{M \times M}$, whose u-th column is written $Q_0(j \mid u)$, and $a_0 = M^{-1} \sum_{u=1}^{M} Q_0(u \mid u)$. The shared checkpoint p_0 is

$$p_0(\cdot \mid t, s, u, \perp) = \begin{cases} \delta_{C_{t+1,u}}, & t < K-1, \\ \sum_j Q_0(j \mid u)\delta_{A_j}, & t = K-1, \end{cases} \qquad p_0(\cdot \mid t, s, u, r) = \delta_{A_r} \quad (r \in [M]).$$

Here δ_v denotes the point mass at token v. The teacher is a frozen copy of this checkpoint, and the initial student uses exactly the same columns. The teacher reference is $r_B = Y$ for $B = 1$ and $r_B = \perp$ for $B = 0$; the student reference is always $\perp$. Write $q_{0,B}(\cdot \mid X, h) = p_0(\cdot \mid t, s, f_{t+1}(s), r_B)$ for its label on a student prefix h with current stage/state (t, s).

Population fitting interface. For a student policy $\pi \in \mathcal{P}$, let μ_π be its distribution of X and sampled prefixes until the first answer or a cap of K emitted tokens. Let $\tau \le K$ be the number of prediction positions visited. The unnormalized occupancy of a student tuple z is

$$d_\pi(z) = \mathbb{E}_{\mu_\pi} \sum_{m=0}^{\tau-1} \mathbf{1}\{z_m = z\}, \qquad z_m = (t_m, s_m, f_{t_m+1}(s_m), \perp).$$

On this fixed occupancy define the token-summed, full-vocabulary forward-KL objective

$$\mathcal{L}_\pi(p) = \mathbb{E}_{X,h \sim \mu_\pi,\, B}\left[\sum_{m=0}^{\tau-1} D_{\mathrm{KL}}\big(q_{0,B}(\cdot \mid X, h_m) \| p(\cdot \mid z_m)\big)\right], \quad p \in \mathcal{P},$$

where $D_{\mathrm{KL}}(q\|p) = \sum_{v\in\mathcal{V}} q(v)\log(q(v)/p(v))$, with the usual zero/infinite conventions. One B is drawn per example and is reused at every label position. A **self-consistently fitted policy** means

$$\pi \in \operatorname*{argmin}_{p\in\mathcal{P}} \mathcal{L}_\pi(p).$$

Prefix sampling and teacher targets are fixed during minimization, and the occupancy is not differentiated through p. Theorem C.1 conditions on this fixed-point fit rather than optimizer convergence. Zero-occupancy columns are unconstrained, so the proof establishes positive occupancy for the relevant states.

Candidate and evaluation quantities. Define

$$Q_\rho = (1-\rho)Q_0 + \rho I_M, \qquad a_\rho = (1-\rho)a_0 + \rho,$$

where I_M is the identity matrix. A candidate $p_\rho \in \mathcal{P}$ has empty-reference columns

$$p_\rho(\cdot \mid t, s, u, \perp) = \begin{cases} (1-\rho)\delta_{C_{t+1,u}} + \dfrac{\rho}{M}\sum_j \delta_{A_j}, & t < K-1, \\ \sum_j Q_\rho(j \mid u)\delta_{A_j}, & t = K-1, \end{cases}$$

and frozen nonempty-reference columns p_0. This candidate is defined on every literal current state/stage, including states inconsistent with earlier tokens. We prove below that it is self-consistently fitted and derive its accuracy.

Free evaluation samples tokens, stops at the first A_j or the K-output cap, and gives reward 1 exactly when that first answer is A_Y. A wrong intermediate value or stage is not itself failure; no answer by the cap is failure. All continuations have the literal interpretation above, so the policy is defined even after such errors. The atomic vocabulary excludes syntactically invalid tokens. Denote expected reward by $A_{\mathrm{free}}(p)$. Teacher accuracy $A_T(p_0;\rho)$ uses the teacher's own sampled prefixes and reference r_B, with the same reward convention. Exact forced evaluation of a closure student suppresses and renormalizes away all answer mass whenever its current stage satisfies $t < K-1$, then samples the ordinary terminal distribution; let $A_{\mathrm{forced}}(p)$ be its expected reward. This supplies no true intermediate states. The theorem's forced evaluation is for p_ρ, whose only continuations follow increasing correct stages. Let $P_{\mathrm{complete}}(p)$ denote the probability of emitting all $K-1$ correct continuations before an answer in free evaluation.

Finite candidate extension. For $0 < \varepsilon < 1$, put $\nu(v) = 1/(KM)$ for every $v \in \mathcal{V}$. On every tuple, define $p_{0,\varepsilon} = (1-\varepsilon)p_0 + \varepsilon\nu$. Teacher and initial student have the identical finite columns $\log p_{0,\varepsilon}(v \mid z)$. Define candidate $p_{\rho,\varepsilon}$ by the same smoothing of p_ρ on empty-reference tuples and by $p_{0,\varepsilon}$ on nonempty-reference tuples. All output tokens remain available in this finite model. Its free reward includes any later correct answer after a smoothing-induced erroneous continuation. Let $A_T(p_{0,\varepsilon};\rho)$ use the smoothed teacher's own rollout and let $A_{\mathrm{free}}(p_{\rho,\varepsilon})$ use the smoothed candidate's rollout, both capped at K. The smoothing is used only for the finite realization and reward-transfer bounds.

B.3.1 Assumptions

Assumption 1 (Fresh permutation task)**.** *$K \geq 2$, $M \geq 2$; independently for each example, $f_1, \ldots, f_K$ are independent uniform permutations of $[M]$, and S_0 is uniform on $[M]$, independent of them. The prompt exposes all tables using the primitive encoding specified above.*

Assumption 2 (Example-level privileged reference)**.** *$0 \leq \rho < 1$; $B \sim \mathrm{Bernoulli}(\rho)$ is independent of X and of all student sampling randomness. The same B labels the whole example. Only the teacher receives r_B.*

Assumption 3 (General checkpoint confusion)**.** *Q_0 is any column-stochastic matrix with $a_0 > 1/M$; zero entries and $a_0 = 1$ are permitted. The no-reference and privileged checkpoint columns are exactly those specified for p_0.*

Assumption 4 (Primitive finite decoder and readout)**.** *The token embeddings, query/key/value projections, causal mask, paired-ReLU weights, residual blocks, and readout restrictions are exactly the construction specified above, with finite $\lambda \geq \log(7(N_{\max}-1))$. Retrieval and decoder weights and all raw-coordinate coefficients of U remain frozen. Only allowed empty-reference output columns are fitted; nonempty-reference columns remain the initial ones. The proof derives retrieval, tuple decoding, and target realization from these fixed weights and restrictions.*

Assumption 5 (Stopped population objective)**.** *Teacher labels always come from the frozen initial checkpoint; student prefixes are generated by the evaluated student; fitting uses precisely $\mathcal{L}_\pi$ with token summation, full-vocabulary forward KL, and stopped occupancy/teacher targets. The result is conditional on a self-consistently fitted policy; it does not assert that a particular optimizer reaches one.*

Assumption 6 (Optional strict-separation subcase)**.** *The additional parameter condition $[1 + (K - 1)/M]/K < a_0 < 1$ is imposed only for the strict monotonicity subclaim. It is not required for the main exact laws, coverage, zero-disclosure recovery, or finite-candidate bounds.*

B.3.2 Fitting and evaluation

The theorem first identifies the self-consistent closure fit and its sampled execution law, then gives a finite smoothed realization with bounded-horizon reward transfer.

B.3.3 Preliminaries

The matrix $Q_\rho = (1-\rho)Q_0 + \rho I_M$ records the terminal answer distribution, and $a_\rho = (1-\rho)a_0 + \rho$ is its uniform diagonal average. Both are defined from the checkpoint and disclosure parameter and enter the execution laws below. The notation $d_\pi(t, s, u, \perp)$ is the expected number of visits to precisely that student tuple, with the token-summed normalization in the setup. A self-consistently fitted policy is always understood with that occupancy frozen in the minimization. All accuracy quantities are expectations of the stated first-answer $0/1$ reward over fresh tasks and sampled tokens, up to K emissions. We use exact probability columns in the closure and finite log-probability columns only for the explicitly positive smoothed policies.

B.4 Theorem 2 setting

B.4.1 Finite composition model

Fix integers $M, K \geq 2$, put $d = K - 1$, and write $[M] = \{1, \ldots, M\}$, $u_0 = 1/M$. A question is $X = (f_1, \ldots, f_K, S_0)$, with states $S_\ell = f_\ell(S_{\ell-1})$ and answer $Y = S_K$. The distribution of questions is specified in Assumption 7. The complete tables are student input, so $H(Y \mid X) = 0$. Each question has a reference $r \in \mathcal{R} = [M] \cup \{\perp\}$. The student always uses $r = \perp$; the teacher uses $r = Y$ if its independent privilege bit B is one and $r = \perp$ otherwise.

The output vocabulary is

$$\mathcal{V} = \{C_{j,i} : j \in [d],\ i \in [M]\} \cup \{A_i : i \in [M]\}, \qquad |\mathcal{V}| = KM.$$

A prompt contains the KM table records in fixed lexicographic order, one reference slot, and a final initial-query token encoding $(t, s) = (0, S_0)$. A continuation token $C_{j,i}$ encodes literally $(t, s) = (j, i)$. At every prediction position the current token supplies (t, s), the next queried table index is $t + 1$, and the retrieved state is $u = f_{t+1}(s)$. Thus an incorrectly ordered or incorrect-state token still defines the next query; it is never replaced by the true state. An answer terminates generation immediately. There are at most K emitted output positions, and no-answer-at-cap is failure. Let h denote a prediction prefix including X and its emitted tokens, and $m \in \{1, \ldots, K\}$ its next emitted position. The stage t in the current token can differ from $m - 1$.

The model is one causal two-head Transformer layer with residual connections, a frozen one-hidden-layer ReLU MLP, and no LayerNorm. Its embeddings use disjoint blocks for table keys, table values, current-token stage/state, reference key/value, head outputs, and a fresh decoded-feature residual block. At prediction positions the step query is (e_{t+1}, e_s); table record $(\ell, i, f_\ell(i))$ has key (e_ℓ, e_i) and value $e_{f_\ell(i)}$. Non-table step keys and values are zero. The effective step score is $\beta(\mathbf{1}_{\ell=t+1} + \mathbf{1}_{i=s})$; usual attention normalization constants can be absorbed into fixed query/key matrices. The reference query is constant, the reference-slot key is one, all other reference keys are zero, its value is e_r, and other reference values are zero, giving effective reference-slot score β. Both heads use ordinary finite softmax over the causal prefix. Their outputs are denoted a^{step} and a^{ref}. The current stage/state blocks remain e_t, e_s, because head outputs occupy separate blocks. Set $N_{\mathrm{ctx}} = KM + K + 2$, a bound on attended context length before any capped prediction.

For each $\zeta = (t', s', u', r') \in \{0, \ldots, d\} \times [M]^2 \times \mathcal{R}$, the frozen MLP writes

$$v_\zeta = (e_t)_{t'} + (e_s)_{s'} + a^{\mathrm{step}}_{u'} + a^{\mathrm{ref}}_{r'}, \qquad \phi_\zeta = 2\{\mathrm{ReLU}(v_\zeta - 13/4) - \mathrm{ReLU}(v_\zeta - 15/4)\}.$$

Its hidden width is $2KM^2(M+1)$, and its output is written into the initially zero feature block. The readout matrix uses only this block; all its coefficients on raw embeddings, residual blocks, and head outputs are fixed at zero. These weights define the finite neural map. Exact selection of $\zeta = (t, s, f_{t+1}(s), r)$ from the finite-softmax heads is proved below.

Write $\sigma(x) = (1+e^{-x})^{-1}$, $\mathrm{logit}(p) = \log(p/(1-p))$, and let the trainable tied coordinates be $\theta = (z_1, \dots, z_d, w) \in \mathbb{R}^K$. Each feature column of the readout matrix has the following coefficients. For a feature with empty reference and $t < d$, put $j = t+1$: the coefficient for $C_{j,u}$ is $z_j + \log M$, for every A_i is zero, and for every other continuation token is $-L$. For empty reference and $t = d$, the coefficient for A_u is $w + \log(M-1)$, for other answers is zero, and for every continuation token is $-L$.

For a disclosed-reference feature with value $r \in [M]$, put

$$r_A = \log \frac{(M-1)a_{\mathrm{ref}}}{1-a_{\mathrm{ref}}}, \qquad r_C = \mathrm{logit}(c_{\mathrm{ref}}) + \log \frac{M-1}{1-a_{\mathrm{ref}}}.$$

The coefficient for A_r is r_A, for other answers is zero; when $t < d$, that for $C_{t+1,u}$ is r_C and for other continuations is $-L$; when $t = d$, all continuation coefficients are $-L$. All these reference coefficients are frozen. Output distributions are the full-vocabulary softmax of the readout logits and are denoted $p_{\theta,L}(\cdot \mid h, r)$, suppressing fixed architecture parameters. No continuation or answer token is masked in training or free evaluation.

The initial tied vector is

$$\theta_0 = (\mathrm{logit}(c_0), \dots, \mathrm{logit}(c_0), \mathrm{logit}(a_0)).$$

The teacher is an exact frozen copy of the entire initial network, including its finite L, β and fixed coefficients. Its label law is $q_{T,L}(\cdot \mid h, B) = p_{\theta_0,L}(\cdot \mid h, r_B)$, where $r_B = Y$ for $B = 1$ and $r_B = \perp$ otherwise. The numbers $c_0, a_0, c_{\mathrm{ref}}, a_{\mathrm{ref}}$ are nominal probabilities after deleting suppressed continuation coefficients, not exact probabilities of the unmasked finite network. In particular the nominal empty-reference terminal answer matrix is symmetric: $Q_0(u \mid u) = a_0$ and $Q_0(i \mid u) = (1-a_0)/(M-1)$ for $i \neq u$.

At iteration n, sample a capped empty-reference student trajectory τ from $p_{\theta_n,L}$; denote its visited prediction prefixes by $\mathcal{H}(\tau)$, of cardinality at most K. Draw one $B \sim \mathrm{Bernoulli}(\rho)$, independently of X, τ, and share it across all labels on this trajectory. The deterministic population semi-gradient algorithm is

$$\theta_{n+1} = \theta_n - \eta\, \mathbb{E}_{X,\tau \sim p_{\theta_n,L},B} \left[\sum_{h \in \mathcal{H}(\tau)} \nabla_\theta \,\mathrm{KL}(q_{T,L}(\cdot \mid h, B) \,\|\, p_{\theta,L}(\cdot \mid h, \perp))|_{\theta=\theta_n} \right].$$

This gradient is Euclidean in the K tied coordinates; it is not the update obtained by independently training duplicated readout entries. The teacher, trajectory sampling law, and prefix contents are held fixed during differentiation. The loss is token-summed, with no trajectory-length normalization. The expectation includes all incorrect-state and out-of-order prefixes reached before stopping or cap. It is an exact population expectation, not a finite sampled batch.

Define $F_L(\theta)$ as free sampled accuracy: the probability that the first answer emitted within the K-token cap is A_Y. A correct final answer after an incorrect intermediate token is success. Define $G_L(\theta)$ as forced accuracy under the following intervention on the same network: at emitted positions $m \le d$, set all answer probabilities to zero and renormalize over all continuation tokens; at $m = K$, use the ordinary unmasked distribution. This rule depends on emitted position, not on the current token's stage. No correct state is supplied; a continuation at the final position is failure. Define $T_L(r) = \Pr(\text{correct capped free answer from the frozen teacher with reference } r)$, where $r = Y$ means the reference varies with the sampled question; thus $T_L(\perp) = F_L(\theta_0)$.

For analysis, define a comparison model by deleting from each feature-column support the structurally suppressed continuation rows: the other continuation rows when $t < d$, and all continuation rows when $t = d$. These are precisely the rows whose coefficients are fixed to $-L$; all active rows and tied parameters are unchanged. This comparison model is used only in the analysis. Apply the same population semi-gradient rule to this model to define $\bar\theta_n = (\bar z_1(n), \dots, \bar z_d(n), \bar w(n))$, initialized at $\bar\theta_0 = \theta_0$, and write $c_j(n) = \sigma(\bar z_j(n))$, $a(n) = \sigma(\bar w(n))$. Denote its free and forced

accuracies by $\bar{F}_n, \bar{G}_n$. The adaptive terminal-visit clock to be derived from its occupancy is

$$H_N = \eta \sum_{n=0}^{N-1} \prod_{j=1}^{d} c_j(n).$$

With $\Delta c = c_0 - c_{\mathrm{ref}}$, $\Delta a = a_{\mathrm{ref}} - a_0$, define

$$c_* = c_0 - \rho\Delta c, \quad a_* = a_0 + \rho\Delta a,$$
$$A_0 = u_0 + (a_0 - u_0)c_0^d, \quad A_* = u_0 + (a_* - u_0)c_*^d, \quad D = A_0 - A_*,$$
$$\lambda_c = \min\{c_0(1-c_0), c_*(1-c_*)\}, \quad \lambda_a = \min\{a_0(1-a_0), a_*(1-a_*)\},$$
$$\gamma = \min\{\lambda_c c_*^{d-1}, \lambda_a c_*^d\}, \quad B_{\mathrm{err}} = \rho\Delta a + d\rho\Delta c.$$

For finite iteration count N, put

$$R_N = \max\{|\operatorname{logit}(c_0)|, |\operatorname{logit}(a_0)|\} + \eta K N,$$
$$E_N(L) = 20K^4 M(N+1)e^{R_N}(1 + 4\eta K^3)^N e^{-L}.$$

Define the final-iterate discrepancy $\delta_{\theta,N} = \|\theta_N - \bar{\theta}_N\|_\infty$, using the maximum absolute difference of the K tied coordinates, and accuracy discrepancies $\delta_{F,n} = |F_L(\theta_n) - \bar{F}_n|$, $\delta_{G,n} = |G_L(\theta_n) - \bar{G}_n|$ for $n \in \{0, N\}$. Theorem C.2 bounds these discrepancies by $E_N(L)$ and relates the comparison dynamics to the finite network.

B.4.2 Assumptions

Assumption 7 (Fresh permutation questions)**.** *Each question independently draws $f_1, \ldots, f_K$ independently and uniformly from the permutation group on $[M]$, and independently draws S_0 uniformly from $[M]$. The privilege bit is independent Bernoulli ρ, shared once per question's rollout. Student sampling is independent of that bit.*

Assumption 8 (Fixed retrieval and tied readout)**.** *Use exactly the disjoint embedding blocks, two softmax heads, paired-ReLU MLP, residual placement, and feature-column readout coefficients above, with finite $\beta \geq \log(7(N_{\mathrm{ctx}} - 1))$ and finite $L > 0$. Only $z_1, \ldots, z_d, w$ are trainable; use the stated literal continuation-token semantics on every visited prefix. The proof derives the decoded features and the resulting dynamics from these weights.*

Assumption 9 (Common finite checkpoint)**.** $0 < c_{\mathrm{ref}} < c_0 < 1$ *and* $u_0 < a_0 < a_{\mathrm{ref}} < 1$. *The teacher and student initially share exactly the entire finite network just specified. Symmetric nominal Q_0 is imposed by these tied coefficients, not by an assumption on unconstrained SGD.*

Assumption 10 (Exact stopped-sampling iteration)**.** *Use exactly the displayed population semi-gradient rule with finite $0 < \eta \leq 1$, full-vocabulary forward KL, stopped teacher and sampling differentiation, and a K-output cap. Free and forced evaluations have precisely the definitions above.*

Assumption 11 (Primitive gain–loss inequality)**.** *For the strict-separation part take $0 < \rho < 1$ and require $\Delta a\, c_0 < d\, \Delta c\, (a_0 - u_0)$. The $\rho = 0$ baseline clause uses the same finite construction and update, without a strict-separation budget.*

Assumption 12 (Explicit numerical budgets)**.** *In the strict-separation part choose a finite integer*

$$N \geq \left\lceil \frac{\log\big(\max\{4, 4B_{\mathrm{err}}/D\}\big)}{\eta\gamma} \right\rceil, \qquad E_N(L) \leq \frac{\min\{D, \rho\Delta a\}}{8}.$$

Assumption 11 implies $D > 0$, after which the proof chooses finite budgets from this positive gap. The inequality is purely numerical and does not assume policy, parameter, or gradient closeness.

B.4.3 Training and evaluation

We prove a finite-iteration execution-risk separation for this finite neural construction and exact population algorithm: forced execution improves, free execution deteriorates, and the same frozen teacher has greater free accuracy when its answer reference is disclosed. The theorem derives

the finite-neural and optimization bridges rather than positing a fitted limiting policy. It separately proves exact checkpoint recovery at zero privilege for every finite iteration, without a strict-separation budget.

The result concerns the tied symmetric readout and exact population iteration; it does not cover sampled SGD, arbitrary Q_0, unrestricted readouts, or practical weight sizes. The ideal execution law appears only in the comparison limit $c_0 \to 1$, $c_{\mathrm{ref}} \to 0$, $a_{\mathrm{ref}} \to 1$, and no finite-iteration rate is uniform in these limits.

B.4.4 PRELIMINARIES

The sigmoid and logit defined in the setup convert the original tied readout coordinates to the comparison continuation probabilities $c_j(n)$ and terminal answer probability $a(n)$. The clock H_N sums comparison terminal visits weighted by η; it measures the training exposure available to the terminal coordinate. The theorem uses these setting quantities to state both the exact iteration and its optimization control. For any finite n, the notation H_n means the same defining sum with horizon n, including $H_0 = 0$.

The values A_0, A_* are the setup's nominal checkpoint and target free accuracies; $D = A_0 - A_*$ measures their signed gap. The endpoint discrepancies $\delta_{F,n}$ and $\delta_{G,n}$ compare the original sampled-success probabilities, while $\delta_{\theta,N}$ compares the same K tied coordinates. For a vector $x \in \mathbb{R}^K$, $\|x\|_\infty = \max_{1 \le k \le K} |x_k|$. These are the metrics used below. The explicit $R_N, E_N(L), \lambda_c, \lambda_a, \gamma, B_{\mathrm{err}}$ retain their setup definitions; none denotes an assumed stability property.

B.5 EXPLICIT FINITE-HORIZON BOUNDS

For Theorem 2, define

$$\begin{aligned}
\lambda_c &= \min\{c_0(1-c_0), c_*(1-c_*)\},\\
\gamma &= \min\{\lambda_c c_*^{d-1}, \lambda_a c_*^d\}, \qquad B_{\mathrm{err}} = \rho\Delta_a + d\rho\Delta_c,\\
R_N &= \max\{|\operatorname{logit} c_0|, |\operatorname{logit} a_0|\} + \eta K N,\\
E_N(L) &= 20K^4 M(N+1) e^{R_N} (1 + 4\eta K^3)^N e^{-L}.
\end{aligned}$$

The finite-training coordinate and endpoint probability errors obey

$$\|\theta_N - \bar\theta_N\|_\infty \le E_N(L), \qquad \max_{k \in \{0,N\}} \max\{|F_L(\theta_k) - \bar F_k|, |G_L(\theta_k) - \bar G_k|\} \le E_N(L).$$

A sufficient choice is

$$\begin{aligned}
N &\ge \left\lceil \frac{\log \max\{4, 4B_{\mathrm{err}}/D\}}{\eta\gamma} \right\rceil,\\
L &\ge \max\left\{1, R_N + N\log(1 + 4\eta K^3) + \log \frac{160K^4 M(N+1)}{\min\{D, \rho\Delta_a\}}\right\}.
\end{aligned}$$

These give $E_N(L) \le \min\{D, \rho\Delta_a\}/8$ and the signed margins in Theorem 2. They are finite-horizon sufficient conditions; no uniform-in-time or singular-limit approximation is asserted. At $\rho = 0$, the checkpoint is exact at every finite iteration; the full argument appears in the proof appendix for Theorem 2.

B.6 THEOREM 3 SETTING

We compare four population distillation protocols within an explicitly constructed neural decoder for permutation composition. The contribution is the exact comparison of their generated policies and sampled accuracies from one shared teacher/student checkpoint. The relative-entropy projection and selection-posterior calculations used below are elementary algebra; the neural realization, collection experiments, and their joint consequences are the subject of the theorem.

Let M, K be integers, $n = K - 1$, $[M] = \{1, \dots, M\}$, and $u = 1/M$. A task is $X = (f_1, \dots, f_K, S_0)$, where each $f_\ell : [M] \to [M]$ is a permutation. Define $S_\ell = f_\ell(S_{\ell-1})$ and $Y = S_K$.

The complete tables of all permutations and S_0 occur in both teacher and student prompts; thus Y is a deterministic function of X. A reference slot contains either $R = \perp$ or $R = Y$. The disclosure indicator is $B = \mathbf{1}\{R \neq \perp\}$. A student always uses $R = \perp$.

The output vocabulary is

$$\mathcal{V} = \{C_{j,i} : 1 \le j \le n, i \in [M]\} \cup \{A_i : i \in [M]\}, \qquad |\mathcal{V}| = KM.$$

Here $C_{j,i}$ literally supplies stage j and state i for the next forward pass. At the initial query the stage and current state are $t = 0, s = S_0$; following any continuation token $C_{j,i}$, they are $t = j, s = i$. Define the next-lookup value $z = f_{t+1}(s)$. For $t < n$, the prescribed continuation token is $C_{t+1,z}$; for $t = n$, this lookup is terminal. These definitions apply even to an incorrect or nonconsecutive state token: no true state replaces the emitted token. Categorical generation stops at its first A_i, or after K emitted tokens if no answer has appeared. A run succeeds exactly when its first answer is A_Y within this cap. A wrong intermediate token is not itself defined as failure. Let $\mathcal{A}(p)$ denote this success probability under a student policy p and fresh tasks. For the initial no-reference policy p_0, write $\mathcal{A}_0 = \mathcal{A}(p_0)$.

The candidate neural family is a one-layer, two-head causal decoder with residual connections, a positionwise ReLU MLP, and no LayerNorm. Let x be the embedding at the last token/query position. Write its last-position computation as

$$h = x + W_O[\mathrm{Attn}_{\mathrm{step}}; \mathrm{Attn}_{\mathrm{ref}}], \qquad \widetilde{h} = h + W_2\,\mathrm{ReLU}(W_1 h), \qquad p_U(\cdot) = \mathrm{softmax}(U\widetilde{h}).$$

All embeddings, attention maps, W_O, W_1, W_2, and reference-conditioned columns of U are frozen. Only no-reference columns of U may change. A constant embedding coordinate permits the displayed linear maps to encode affine thresholds without changing this architecture. Separate embedding blocks contain table keys, table values, generated query fields, stage fields, and a reference-slot marker. At the last query, the step query is $e_{t+1} \oplus e_s$; table entry (ℓ, i) has key $e_\ell \oplus e_i$ and value $e_{f_\ell(i)}$. Non-table keys/values for this head are zero. The reference head selects the unique reference-slot marker and returns (B, Be_Y), with zero answer field when $B = 0$. Here e_i denotes the unit vector in the indicated field, and $\oplus$ concatenates fields. Both heads read embeddings in parallel; the reference head cannot use the current step-head result as a second dependent lookup.

The proposed frozen MLP has disjoint coordinates indexed by

$$\phi^0_{t,z} = \mathbf{1}\{\text{stage} = t,\ \text{step value} = z,\ B = 0\},$$
$$\phi^1_{t,z,y} = \mathbf{1}\{\text{stage} = t,\ \text{step value} = z,\ B = 1,\ R = y\}.$$

The intended construction uses one-hot conjunctions $\mathrm{ReLU}(\sum_{r=1}^{m} v_r - (m-1))$, where m is the number of conjoined fields and the $v_r \in \{0,1\}$, $1 \le r \le m$, are their indicators. It writes these coordinates into reserved residual slots; U reads only these slots, with coefficients on all raw residual, query, table and head fields fixed at zero. Stage columns are distinct, so fitting need not use a single continuation gate at all stages. The proof below verifies these indicators with explicit frozen weights of polynomial dimension and width in K, M.

Let $Q_b = (Q_b(i \mid y))_{i,y \in [M]}$, for $b \in \{0,1\}$, be confusion matrices. Define

$$a_b = \frac{1}{M}\sum_{y=1}^{M} Q_b(y \mid y), \qquad \overline{Q}_1(i) = \frac{1}{M}\sum_{y=1}^{M} Q_1(i \mid y).$$

With continuation parameters c_0, c_1, the shared checkpoint specification is the following distribution $q_b(\cdot \mid X, \xi)$ on an unanswered prefix ξ, interpreted using its literal (t, s, z):

$$\begin{aligned} t < n: \quad & q_0 = c_0\delta_{C_{t+1,z}} + (1-c_0)\textstyle\sum_i u\delta_{A_i}, \\ & q_1 = c_1\delta_{C_{t+1,z}} + (1-c_1)\textstyle\sum_i Q_1(i \mid Y)\delta_{A_i}; \\ t = n: \quad & q_0 = \textstyle\sum_i Q_0(i \mid z)\delta_{A_i}, \qquad q_1 = \textstyle\sum_i Q_1(i \mid Y)\delta_{A_i}. \end{aligned}$$

The symbol δ_v denotes unit mass on token v. The teacher is the frozen checkpoint θ_0, evaluated with the appropriate reference. All four student copies begin at that exact checkpoint and therefore initially have $p_0 = q_0$, including its continuation and early stopping behavior. Reference columns remain frozen at their initial values.

Exact zero probabilities are interpreted in the closure of the readout softmax family. For a target column $v \in \Delta(\mathcal{V})$, where $\Delta(\mathcal{V})$ is the probability simplex, the positive distribution $v_\varepsilon = (1-\varepsilon)v + \varepsilon\,\mathrm{Unif}(\mathcal{V})$, $0 < \varepsilon < 1$, is represented by logits $\log v_\varepsilon$ and converges to v. Denote by Π_0 the resulting closure with frozen reference columns and separately variable no-reference feature columns. The relevant student readout feature is $F = (t, z)$. The construction below verifies this feature map and the independent simplex columns in Π_0. Theorem C.3 is stated for this hard-attention closure.

The four collection/labeling protocols are indexed by $P \in \{\mathrm{ON}, \mathrm{RETAIN}, \mathrm{RELABEL}, \mathrm{PROCESS}\}$. A collected trajectory contributes every unanswered prefix before a sampled output, including the prefix at which its first answer is sampled, up to the K-token cap. Its label q is a full distribution on $\mathcal{V}$, not a sampled teacher token. The joint experiment is:

Protocol P	Prefix collection	Label distribution at a collected prefix
ON	Candidate fitted student p_P, empty reference	q_B, where one independent $B \sim \mathrm{Bernoulli}(\rho)$ is drawn per task and reused at every label
RETAIN	Frozen teacher $q_{B_{\mathrm{col}}}$, where $B_{\mathrm{col}} \sim \mathrm{Bernoulli}(\rho)$ is drawn once	$q_{B_{\mathrm{col}}}$, using the original collection bit at every label
RELABEL	Frozen teacher $q_{B_{\mathrm{col}}}$, with the same collection rule as RETAIN	$q_{B_{\mathrm{lab}}}$, with a new independent $B_{\mathrm{lab}} \sim \mathrm{Bernoulli}(\rho)$ drawn once and reused at every label
PROCESS	Candidate fitted student p_P, empty reference	q_0 at $t < n$, including the checkpoint's premature answers; q_B at $t = n$, with one independent $B \sim \mathrm{Bernoulli}(\rho)$ per task

All disclosure bits are independent of the task; in RELABEL the labeling bit is also independent of the collection bit and collection randomness. Student decoding randomness is independent of its labeling bit. Teacher collection uses only the original collection bit. Conditional on the task, applicable bits and past tokens, each token is drawn from the specified categorical policy with fresh randomness. PROCESS changes the supervision rule; RELABEL changes the joint collection/labeling experiment. Neither protocol provides correct intermediate tokens as student inputs.

For a fixed collection law μ, define the token-summed population objective

$$J_P(r;\mu) = \mathbb{E}_\mu\left[\sum_{\xi\ \mathrm{collected}} D_{\mathrm{KL}}\big(q_P(\cdot \mid X, \xi)\,\|\,r(\cdot \mid F(\xi))\big)\right], \qquad r \in \Pi_0.$$

Here $D_{\mathrm{KL}}(q\|r) = \sum_{v\in\mathcal{V}} q(v)\log(q(v)/r(v))$, with $0\log(0/r) = 0$ and positive mass against zero giving $+\infty$; q_P is the random label prescribed by the protocol. The sum is not divided by trajectory length. The collection law, prefix sampling, and labels are held fixed in this minimization or any associated student derivative. Write $\mu_P[p]$ for the ON or PROCESS joint experiment collected under p; μ_P without an argument denotes the fixed teacher-collected law for RETAIN or RELABEL. Positive occupancy of a feature f means $\mathbb{E}_\mu \sum_{\xi\ \mathrm{collected}} \mathbf{1}\{F(\xi) = f\} > 0$. Fitting on a zero-occupancy feature is not prescribed. Define $\mathcal{A}_P = \mathcal{A}(p_P)$ for a fitted policy, and let $\mathcal{A}_T(\rho)$ be frozen-teacher accuracy when the per-task reference bit has parameter ρ.

Assumption 13 (Fresh permutation tasks)**.** *$K \geq 2$, $M \geq 2$, the f_ℓ are independent uniform elements of the permutation group on $[M]$, and S_0 is independent and uniform on $[M]$. Tasks are independent across examples. The proof derives the required conditional laws from this task distribution.*

Assumption 14 (Controlled checkpoint parameters)**.** *$0 \leq c_1 < c_0 \leq 1$; for $b = 0, 1$, $Q_b(i \mid y) \geq 0$ and $\sum_i Q_b(i \mid y) = 1$ for every y; and $u < a_0 \leq a_1 \leq 1$. The displayed checkpoint column specification uses these same constants at every state and unfinished stage, with one shared initial teacher/student checkpoint and frozen reference columns. Symmetry and double stochasticity of either matrix are not required. The compatible neural checkpoint is constructed explicitly below.*

Assumption 15 (Restricted neural regime)**.** *Attention uses the unique-maximum hard-attention operation; input/head/MLP routing and the frozen-zero raw-feature readout restrictions are exactly*

those specified above. Readout distributions and population minimization are taken in the closure Π_0*. This assumption fixes the architecture and parameter domain used by Theorem C.3.*

Assumption 16 (Disclosures and collection protocol)**.** $0 \le \rho < 1$*, and each arm follows its specified joint task, once-per-example bit, collection, labeling and sampling experiment. No bit is redrawn per token; the selection posterior and accuracy laws are derived below.*

Assumption 17 (Conditional population fitting)**.** *For each teacher-collected arm* $P \in \{\mathrm{RETAIN}, \mathrm{RELABEL}\}$*, the policy being compared satisfies* $p_P \in \arg\min_{r\in\Pi_0} J_P(r;\mu_P)$*. For* $P \in \{\mathrm{ON}, \mathrm{PROCESS}\}$*, it satisfies the self-consistent condition* $p_P \in \arg\min_{r\in\Pi_0} J_P(r;\mu_P[p_P])$*, where* $\mu_P[p_P]$ *is frozen during minimization over* r*. These are the only optimization conditions; the proof derives the conditional averages and fitted-policy existence, without asserting convergence of finite gradient updates.*

Under these assumptions, the proof constructs the shared checkpoint and fitted policies and derives their occupancies, conditional targets, and sampled accuracies.

B.6.1 Preliminaries

For $0 \le j \le n$, let E_j be the event that the first j outputs are the correct continuation tokens $C_{1,S_1}, \dots, C_{j,S_j}$, with E_0 the sure event; at that next prediction define $F_j = (j, Z_j)$, $Z_j = S_{j+1}$. Define the scalar expressions

$$\bar{c} = (1-\rho)c_0 + \rho c_1, \qquad a_{\mathrm{mix}} = (1-\rho)a_0 + \rho a_1, \qquad D_j = (1-\rho)c_0^j + \rho c_1^j, \quad w_j = \frac{\rho c_1^j}{D_j},$$

using $c_b^0 = 1$, including $0^0 = 1$ solely for these depth-zero expressions. Set

$$Q_{\mathrm{mix}} = (1-\rho)Q_0 + \rho Q_1, \qquad Q_{\mathrm{ret},n} = (1-w_n)Q_0 + w_n Q_1.$$

For recording the target projections, define candidate label-bit weights $h_{P,j}$ by $h_{\mathrm{ON},j} = h_{\mathrm{RELABEL},j} = \rho$, $h_{\mathrm{RETAIN},j} = w_j$, and $h_{\mathrm{PROCESS},j} = 0$ for $j < n$, $h_{\mathrm{PROCESS},n} = \rho$. These are candidate values to derive, not stipulated conditional probabilities. Let I_M denote the identity confusion matrix, $I_M(i \mid y) = \mathbf{1}\{i = y\}$. For the source specialization define the threshold

$$T_{K,M}(\rho) = u + \frac{(1-u)\rho(1-\rho)^{K-2}}{1-(1-\rho)^{K-1}}, \qquad 0 < \rho < 1.$$

The quantities D_j, w_j express the candidate survival mixture and original-bit weight; $\bar{c}, a_{\mathrm{mix}}, Q_{\mathrm{mix}}$ express the independent-label mixtures. They are defined from the original checkpoint and disclosure parameters solely to state the exact columns, accuracy identities, and source comparison below. Their asserted probabilistic interpretations and positivity are conclusions of the proof.

C Proofs of the main theorems

For each main result, we restate the full technical claim and then prove it using the supporting results in Appendix D.

C.1 Theorem 1: terminal prediction and survival

Theorem C.1 (Constructive conditional population separation)**.** *Under Assumptions 1, 2, 3, 4, and 5, the prescribed one-layer, two-head network satisfies*

$$\phi = e_{(t,s,f_{t+1}(s),r)}$$

at every allowed literal prediction within the K*-emission cap, including after incorrect continuations. Its hidden and decoded widths are exactly* $2KM^2(M+1)$ *and* $KM^2(M+1)$*; its primitive/head width is* $O(K+M)$*, with only an absolute coordinate-bookkeeping constant. The shared readout selects independent tuple columns and excludes every raw coordinate.*

The candidate p_ρ *is self-consistently fitted. If* $\pi \in \mathcal{P}$ *is any self-consistently fitted policy, then for all* $0 \le t \le K-1$ *and* $s, u \in [M]$*,*

$$d_\pi(t,s,u,\perp) = \frac{(1-\rho)^t}{M^2} > 0, \qquad \pi(\cdot \mid t,s,u,\perp) = p_\rho(\cdot \mid t,s,u,\perp).$$

For any fixed closure policy, surviving prefixes are correct. At every occupied unfinished student tuple, the unread suffix retains its independent permutation law, Y is uniform on $[M]$, and the example-level bit retains its Bernoulli law independently of Y; at terminal tuples $Y = u$. The stopped forward-KL minimizer is uniquely the conditional teacher mean at each occupied column; null columns are not identified by the loss.

The exact sampled laws are

$$A_T(p_0;\rho) = A_{\text{forced}}(p_\rho) = a_\rho, \qquad P_{\text{complete}}(p_\rho) = (1-\rho)^{K-1},$$

$$A_{\text{free}}(p_\rho) = \frac{1}{M} + \left(a_\rho - \frac{1}{M}\right)(1-\rho)^{K-1}.$$

Every fitted π has the same student laws. At $\rho = 0$, $p_{\rho=0} = p_0$ on every student tuple and $A_{\text{free}}(p_{\rho=0}) = a_0$. If additionally Assumption 6 holds, then, with K, M, Q_0 fixed, teacher and forced accuracies strictly increase and fitted free accuracy strictly decreases on $[0,1)$.

Separately, for every $0 < \varepsilon < 1$, the declared smoothed teacher, identical initial student, and smoothed candidate admit finite unmasked-softmax log columns with

$$|\log p(v \mid z)| \le \log(KM/\varepsilon) \quad \textit{for every tuple and vocabulary token},$$

and their actual rewards satisfy

$$|A_{\text{free}}(p_{\rho,\varepsilon}) - A_{\text{free}}(p_\rho)| \le 1 - (1-\varepsilon)^K \le K\varepsilon,$$
$$|A_T(p_{0,\varepsilon};\rho) - A_T(p_0;\rho)| \le 1 - (1-\varepsilon)^K \le K\varepsilon.$$

At zero disclosure the exact finite identity is $p_{\rho=0,\varepsilon} = p_{0,\varepsilon}$ on all tuples. The finite rewards include later successful answers after erroneous continuations. The representation is deterministic within the cap, and the reward comparisons are population expectations at this finite horizon.

The population result is conditional on the stated fit, whose existence is proved constructively. The smoothed extension provides a finite realization and reward-transfer bound; it does not add an optimizer-convergence claim.

Proof of Theorem 1. For a frozen collection law, let ν_ζ be the expected number of visits to feature ζ, and let m_ζ be its occupancy-weighted teacher mean. Expanding the cross-entropy gives

$$\mathcal{L}_\mu(p) = C_\mu + \sum_{\zeta:\nu_\zeta>0} \nu_\zeta D_{\text{KL}}(m_\zeta \| p_\zeta), \qquad p^*_\zeta = m_\zeta \quad (\nu_\zeta > 0), \tag{C.1}$$

where C_μ is independent of p. At an unfinished selected feature $\zeta = (t, s, z, \perp)$, the fresh unread suffix leaves Y uniform, even after conditioning on survival. Thus

$$m_\zeta = \begin{cases} (1-\rho)\delta_{C_{t+1,z}} + \rho U_{\mathcal{A}}, & t < d, \\ (1-\rho)\mathsf{Q}_0(z) + \rho\delta_{A_z}, & t = d. \end{cases}$$

Since $1 - \rho > 0$, induction from the initial prompt gives positive occupancy at every relevant stage. These columns also define a self-consistent candidate, so the fit is nonvacuous and every fitted policy has the stated behavior (Propositions D.4 and D.5).

Let E_t denote survival through t correct continuations. Then $\Pr(E_t) = (1-\rho)^t$, and the first-answer events $E_t \setminus E_{t+1}$ partition early stopping. Their success contribution is $u\rho(1-\rho)^t$, giving

$$A_{\text{free}} = u\sum_{t=0}^{d-1} \rho(1-\rho)^t + a_\rho(1-\rho)^d = u\big[1 - (1-\rho)^d\big] + a_\rho(1-\rho)^d. \tag{C.2}$$

Forced renormalization retains the unique correct continuation at each stage and therefore yields a_ρ. Finally,

$$\frac{\partial A_{\text{free}}}{\partial \rho} = (1-\rho)^{d-1}\big[(1-a_0) - d(a_0-u) - (d+1)\rho(1-a_0)\big].$$

The stated threshold makes the bracket negative, while $\partial_\rho a_\rho = 1 - a_0 > 0$. The supporting construction, occupancy, and finite-realization results used above are proved in Section D.1. □

C.2 THEOREM 2: FINITE POPULATION TRAINING

Theorem C.2 (Finite-iteration execution-risk separation)**.** *Under Assumptions 7, 8, 9, and 10, the finite network selects exactly $\phi = e_{(t,s,f_{t+1}(s),r)}$ at every capped prediction prefix, for every reference and finite tied vector, including incorrect-state and out-of-order prefixes. Its logits are exactly the stipulated feature column, with the full output vocabulary retained.*

For $\rho = 0$ or $0 < \rho < 1$, at every finite horizon N, the comparison process obeys, for $0 \le n < N$ and $j \in [d]$,

$$\bar{z}_j(n+1) = \bar{z}_j(n) + \eta \left(\prod_{i<j} c_i(n) \right) (c_* - c_j(n)),$$

$$\bar{w}(n+1) = \bar{w}(n) + \eta \left(\prod_{i=1}^{d} c_i(n) \right) (a_* - a(n)).$$

For $0 \le n \le N$,

$$\bar{F}_n = u_0 + (a(n) - u_0) \prod_{i=1}^{d} c_i(n), \qquad \bar{G}_n = a(n),$$

$$c_* \le c_j(n) \le c_0, \qquad a_0 \le a(n) \le a_*,$$

$$0 \le c_j(n) - c_* \le \rho\Delta c\, e^{-\eta\gamma n}, \quad 0 \le a_* - a(n) \le \rho\Delta a\, e^{-\eta\gamma n},$$

$$a_* - a(n) \le \rho\Delta a\, e^{-\lambda_a H_n}, \qquad |\bar{F}_n - A_*| \le B_{\mathrm{err}} e^{-\eta\gamma n}.$$

Both clock lower bounds hold:

$$H_N \ge \eta \sum_{n=0}^{N-1} \sigma(\mathrm{logit}(c_0) - \eta n \rho \Delta c)^d, \qquad H_N \ge \eta N c_*^d.$$

The actual finite population process satisfies the exact numerical bridge

$$\delta_{\theta,N} \le E_N(L), \qquad \max_{n \in \{0,N\}} \max\{\delta_{F,n}, \delta_{G,n}\} \le E_N(L).$$

In the strict regime of Assumption 11, $D > 0$ and $\gamma > 0$, and finite budgets in Assumption 12 exist. Every such finite N, L gives

$$G_L(\theta_N) - G_L(\theta_0) \ge \frac{\rho\Delta a}{2} > 0, \qquad F_L(\theta_0) - F_L(\theta_N) \ge \frac{D}{2} > 0,$$

$$T_L(Y) > T_L(\perp).$$

Independently, in the $\rho = 0$ baseline clause of Assumption 11, for every finite n and all construction-admissible finite L, β,

$$\theta_n = \theta_0, \qquad F_L(\theta_n) = F_L(\theta_0), \qquad G_L(\theta_n) = G_L(\theta_0).$$

This baseline is exact and uses neither the strict gain–loss inequality nor Assumption 12.

All bounds are deterministic population statements at the fixed rollout cap and finite horizon. With the remaining quantities fixed, $E_N(L) \to 0$ as $L \to \infty$; the theorem does not claim a uniform-in-time approximation.

Corollary C.1 (Explicit finite budgets for the signed margins)**.** *Under Assumptions 7, 8, 9, 10, and the strict clause of Assumption 11, $D, \gamma > 0$. Choose any finite integer*

$$N \ge \left\lceil \frac{\log(\max\{4, 4B_{\mathrm{err}}/D\})}{\eta\gamma} \right\rceil$$

and any finite

$$L \ge \max\left\{1, R_N + N\log(1 + 4\eta K^3) + \log \frac{160K^4M(N+1)}{\min\{D, \rho\Delta a\}}\right\}.$$

Then the forced gain is at least $\rho\Delta a/2$, the free loss is at least $D/2$, and the frozen teacher's disclosed-reference advantage is at least $(1 - \rho/4)\Delta a > 0$. The conventions are those of Theorem C.2.

Proof. Lemma D.26 proves positivity before these choices. Proposition D.29 verifies the budgets. Proposition D.31 transfers the comparison margins to the finite network, and Proposition D.32 gives the teacher advantage. □

Proof of Theorem 2. In the comparison model, differentiating a local continuation loss in its logit gives

$$\frac{\partial}{\partial z} D_{\mathrm{KL}}(\mathrm{Ber}(c_*) \,\|\, \mathrm{Ber}(\sigma(z))) = \sigma(z) - c_*.$$

The probability of visiting stage j is $\prod_{i<j} c_i(n)$. Multiplying the local derivative by this occupancy gives (4.2); the terminal calculation gives (4.3). The sigmoid updates preserve $c_j(n) \in [c_*, c_0]$ and $a(n) \in [a_0, a_*]$. With $e_a(n) = a_* - a(n)$ and $v_n = \prod_j c_j(n)$, the mean value theorem yields

$$\begin{aligned} e_a(n+1) &= \big[1 - \eta v_n \sigma'(\xi_n)\big] e_a(n) \le (1 - \eta \lambda_a v_n) e_a(n), \\ e_a(N) &\le \rho\Delta_a \prod_{n<N} (1 - \eta\lambda_a v_n) \le \rho\Delta_a \exp\left(-\lambda_a \eta \sum_{n<N} v_n\right). \end{aligned} \tag{C.3}$$

The analogous continuation contraction and $v_n \ge c_*^d$ establish the stated limits (Proposition D.21).

To transfer the result, write

$$\bar{G}_n = a(n), \qquad \bar{F}_n = u + [a(n) - u] \prod_j c_j(n), \qquad A_* = u + (a_* - u) c_*^d.$$

Condition (4.5) makes $(a_0 - u + r\Delta_a)(c_0 - r\Delta_c)^d$ strictly decreasing for $0 \le r \le \rho$, hence $D > 0$ (Lemma D.26). Comparison convergence allows a finite N with $\varepsilon_a = a_* - a(N) \le \rho\Delta_a/4$ and $\varepsilon_F = |\bar{F}_N - A_*| \le D/4$. Proposition D.28 bounds each free/forced endpoint discrepancy by $E_N(L) = C_N e^{-L}$, with explicit finite C_N independent of L. Its whole-trajectory coupling includes erroneous continuations and later recovery. Choose finite L so that $E_N(L) \le \min\{D, \rho\Delta_a\}/8$. Accounting for both endpoints gives

$$\begin{aligned} G_L(\theta_N) - G_L(\theta_0) &\ge \rho\Delta_a - \varepsilon_a - 2E_N(L) \ge \tfrac{1}{2}\rho\Delta_a, \\ F_L(\theta_0) - F_L(\theta_N) &\ge D - \varepsilon_F - 2E_N(L) \ge \tfrac{1}{2} D. \end{aligned} \tag{C.4}$$

This proves the finite-network conclusion. The finite-network comparison, endpoint transfer, and explicit-budget results invoked above are proved in Section D.2. □

C.3 Theorem 3: collection and supervision

Theorem C.3 (Constructive conditional protocol comparison)**.** *Under Assumptions 13–17, there is a compatible one-layer, two-head decoder with MLP width $KM(M+1)$ and residual dimension $KM(M+1) + 3K + 6M + 4$. Its frozen embeddings and matrices implement the prescribed literal lookup, reference selection, and disjoint indicators on every literal unanswered input. Its no-reference readout has actual feature $F = (t, z)$ and independently variable simplex columns in Π_0. One common closure checkpoint realizes q_0, q_1, and all four students begin with exactly $p_0 = q_0$.*

Fitted policies exist for all four arms, including the required self-consistent ON and PROCESS fits. Every policy satisfying its corresponding conditional-fit premise has, for all $j < n$, $z \in [M]$, the full column

$$\begin{aligned} p_P(\cdot \mid (j,z)) = {} & [(1 - h_{P,j})c_0 + h_{P,j} c_1]\delta_{C_{j+1,z}} \\ & + \sum_i [(1 - h_{P,j})(1 - c_0)u + h_{P,j}(1 - c_1)\overline{Q}_1(i)]\delta_{A_i}, \end{aligned}$$

and terminal column

$$p_P(\cdot \mid (n,z)) = \sum_i [(1 - h_{P,n}) Q_0(i \mid z) + h_{P,n} Q_1(i \mid z)]\delta_{A_i}.$$

All unlisted coordinates vanish. Thus ON and RELABEL have continuation mass $\bar{c}$ and terminal matrix Q_{mix}; RETAIN has mass D_{j+1}/D_j and terminal matrix $Q_{\mathrm{ret},n}$; PROCESS has mass c_0 and terminal matrix Q_{mix}. The original labeling bit under teacher collection satisfies

$$\Pr(B_{\mathrm{col}} = 1 \mid E_j, F_j = (j,z)) = w_j \qquad (0 \le j \le n).$$

Every correct-depth feature has positive token-count training occupancy: D_j/M for teacher-collected arms, $\bar{c}^j/M$ for ON, and c_0^j/M for PROCESS. Before terminal, Y is uniform conditional on the relevant selected readout feature and surviving generated history; this also holds conditional on $(S_0, f_1, \ldots, f_{j+1})$ and survival, whenever the selected event is positive. At terminal $Y = z$. The ON and RELABEL labeling weights remain ρ; the original teacher bit has weight w_j; PROCESS has effective weight zero before terminal and ρ at terminal.

All evaluated student continuations are correct, consecutive, and self-generated almost surely. They answer by output K; arbitrary zero-probability training prefixes cannot alter their evaluation behavior. Their premature-answer vectors may be nonuniform, but conditional success given an unfinished depth, its readout feature, and stopping is u whenever that stop event is positive; the pooled premature-stop event has the same conditional success when positive. For fresh tasks and categorical first-answer sampling,

$$\begin{aligned}
\mathcal{A}_0 &= u + (a_0 - u)c_0^n, \\
\mathcal{A}_{\mathrm{ON}} = \mathcal{A}_{\mathrm{RELABEL}} &= u + (a_{\mathrm{mix}} - u)\bar{c}^{\,n}, \\
\mathcal{A}_{\mathrm{RETAIN}} &= u + (1-\rho)c_0^n(a_0 - u) + \rho c_1^n(a_1 - u), \\
\mathcal{A}_{\mathrm{PROCESS}} &= u + (a_{\mathrm{mix}} - u)c_0^n, \\
\mathcal{A}_T(\rho) &= (1-\rho)\mathcal{A}_0 + \rho a_1.
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathcal{A}_{\mathrm{PROCESS}} - \mathcal{A}_0 &= \rho(a_1 - a_0)c_0^n \geq 0, \\
\mathcal{A}_{\mathrm{PROCESS}} - \mathcal{A}_{\mathrm{ON}} &= (a_{\mathrm{mix}} - u)(c_0^n - \bar{c}^n) \geq 0, \\
\mathcal{A}_T(\rho) - \mathcal{A}_T(0) &= \rho(a_1 - \mathcal{A}_0) \geq 0.
\end{aligned}$$

At $\rho = 0$, every fitted arm recovers the identical initial checkpoint q_0 on every reachable feature and has the same taskwise stopped trajectory law; all four accuracies equal $\mathcal{A}_0$.

These identities hold in the hard-attention probability-column closure with the first-answer rule and finite K-output cap, under the population-fitting condition of Assumption 17.

Corollary C.2 (Sharp comparison in the source specialization)**.** *Under Assumptions 13–17, specialize to $c_0 = 1, c_1 = 0, Q_1 = I_M$ and $0 < \rho < 1$. Then*

$$\begin{aligned}
&\mathcal{A}_{\mathrm{ON}} - \mathcal{A}_{\mathrm{RETAIN}} = (1-\rho)[1 - (1-\rho)^{K-1}](T_{K,M}(\rho) - a_0), \\
&\mathcal{A}_{\mathrm{ON}} \leq \mathcal{A}_{\mathrm{RETAIN}} \quad \Longleftrightarrow \quad a_0 \geq T_{K,M}(\rho).
\end{aligned}$$

Equality holds exactly at the threshold. For $K = 2$, $T_{2,M}(\rho) = 1$, so ON strictly exceeds RETAIN whenever $a_0 < 1$. For each fixed $K > 2, M \geq 2, 0 < \rho < 1$, both strict orderings and equality occur among admissible Q_0. The formulas are exact sampled first-answer probabilities through K outputs.

Proof. Set $c_0 = 1, c_1 = 0, Q_1 = I_M$, so $a_1 = 1$. Proposition D.43 gives the displayed difference and its positive prefactor. Lemma D.44 gives the threshold range, and Proposition D.44 constructs admissible matrices on both sides of the threshold. □

Proof of Theorem 3. Let E_j be survival through j correct continuations and $F_j = (j, S_{j+1})$ the selected student feature. Constant teacher continuation and the fresh permutation law give the joint masses

$$\Pr(E_j, F_j = (j,z), B_{\mathrm{col}} = b) = \Pr(B_{\mathrm{col}} = b)\frac{c_b^j}{M}, \qquad \Pr(E_j, F_j = (j,z)) = \frac{D_j}{M}. \tag{C.5}$$

Their ratio shows that the posterior is w_j even conditional on the fitted feature, not just on depth. RELABEL's independently drawn labeling bit instead retains posterior ρ (Lemmas D.38 and D.39).

Conditional KL minimization averages the teacher columns. Put $h_{\mathrm{RETAIN},j} = w_j$ and $h_{\mathrm{RELABEL},j} = h_{\mathrm{ON},j} = \rho$, and let $\overline{Q}_1(i) = M^{-1}\sum_y Q_1(i \mid y)$. For $j < d$, the selected uniform law of Y gives

$$\begin{aligned}
p_P(C_{j+1,z} \mid j, z) &= (1 - h_{P,j})c_0 + h_{P,j}c_1, \\
p_P(A_i \mid j, z) &= (1 - h_{P,j})(1 - c_0)u + h_{P,j}(1 - c_1)\overline{Q}_1(i).
\end{aligned} \tag{C.6}$$

At terminal, $Y = z$, so the answer column is $(1 - h_{P,d})Q_0(\cdot \mid z) + h_{P,d}Q_1(\cdot \mid z)$. These are feasible simplex columns. The constructed candidates have positive occupancy; checking their own collection laws also establishes the self-consistent ON fit (Proposition D.38).

Substitution gives (4.8)–(4.9), with RETAIN survival

$$\prod_{j=0}^{d-1} c_{\mathrm{RETAIN},j} = \prod_{j=0}^{d-1} \frac{D_{j+1}}{D_j} = D_d.$$

Although the early-answer column in (C.6) need not be uniform, its contribution to correct early answers is u times its stopping mass, because the unresolved answer remains uniform and the read-out uses only the selected feature. Combining first-answer events with the terminal channel gives the protocol accuracy laws. The fitted-column construction, self-consistency argument, and source-specialization comparison used above are proved in Section D.3. □

D SUPPORTING LEMMAS AND PROPOSITIONS

D.1 THEOREM 1

Throughout this section, we work under the setting in Appendix B.3.

D.1.1 EXACT FINITE TUPLE DECODING AND COLUMN REALIZATION

Lemma D.1 (Finite head margins)**.** *Under the standing assumptions, fix any allowed prediction position before the first answer and within the K-emission cap. Let (t, s) be its literal current stage/state, let r be its reference, and put $u = f_{t+1}(s)$. If n is its number of causally visible tokens, then $n \le N_{\max}$ and*

$$7/8 \le v_u^{\mathrm{step}} \le 1, \qquad 0 \le v_j^{\mathrm{step}} \le 1/8 \quad (j \ne u),$$

$$7/8 \le v_r^{\mathrm{ref}} \le 1, \qquad 0 \le v_j^{\mathrm{ref}} \le 1/8 \quad (j \ne r).$$

Proof. For $0 \le m < K$ previous continuations,

$$n = KM + 2 + m \le KM + K + 1 = N_{\max}.$$

Literal token semantics give $0 \le t < K$, irrespective of prefix correctness. The step-head score of table record (ℓ, i) is

$$\mathbf{1}\{\ell = t + 1\} + \mathbf{1}\{i = s\}.$$

Its unique maximizer is $(t + 1, s)$, with margin at least one; the reference slot likewise has margin one. Thus, for either head,

$$w_* = \frac{1}{1 + \sum_{j \ne *} \exp(\lambda(g_j - g_*))} \ge \frac{1}{1 + (n-1)e^{-\lambda}}.$$

$$(n-1)e^{-\lambda} \le (N_{\max} - 1)e^{-\lambda} \le \frac{1}{7}, \qquad w_* \ge \frac{1}{1 + 1/7} = \frac{7}{8}.$$

The winning values are e_u, e_r; all values are zero or basis vectors. Hence the winning coordinate is at least $7/8$, and every other coordinate is at most $1 - w_* \le 1/8$. □

Lemma D.2 (Exact tuple decoding)**.** *Under the standing assumptions, at every allowed prediction position*

$$\phi = e_{(t,s,f_{t+1}(s),r)}.$$

This computation uses the prescribed one causal layer with exactly two heads and one residual MLP with a single hidden ReLU layer; the hidden width is $2KM^2(M + 1)$, the decoded width is $KM^2(M+1)$, and primitive/head coordinates require only $O(K+M)$, with an absolute bookkeeping constant.

Proof. Put $u = f_{t+1}(s)$. The primitive stage/state blocks remain exact, so

$$b_{(t,s,u,r)} = 2 + v_u^{\text{step}} + v_r^{\text{ref}} \geq 2 + 7/8 + 7/8 = 15/4.$$

For a false tuple with wrong stage or state, $b \leq 1+1+1 = 3$. Otherwise a head coordinate is wrong, giving

$$b_{(t',s',u',r')} \leq 2 + 1/8 + 1 = 25/8 < 13/4.$$

For $b \leq 13/4$ both ReLUs vanish; for $b \geq 15/4$,

$$\text{ReLU}(b - 13/4) - \text{ReLU}(b - 15/4) = (b - 13/4) - (b - 15/4) = 1/2.$$

Thus $\phi_z = 2[\text{ReLU}(b_z - 13/4) - \text{ReLU}(b_z - 15/4)]$ equals one exactly at the true tuple. There are $KM^2(M+1)$ tuples and two hidden units per tuple. The disjoint primitive/head blocks use $O(K+M)$ coordinates; the stage shift is a fixed linear relabeling. All attention inputs are primitive embeddings, so one causal layer suffices. □

Proposition D.1 (Independent column realization)**.** *Under the standing assumptions, at each allowed prediction with tuple z the network's vocabulary logit vector equals the z-column of U's decoded block. These columns may be specified independently subject to the setting's frozen-column restrictions. For every strictly positive collection of probability columns on the vocabulary, its finite log-probability columns realize it. Every collection of probability columns, including the setup's support-restricted closure columns and frozen reference columns, is a simultaneous limit of finite unmasked softmax realizations.*

Proof. Since $\phi = e_z$ and all raw-coordinate readout coefficients vanish,

$$\sum_{z' \in \mathcal{Z}} U_{v,z'} \phi_{z'} = U_{v,z}.$$

Distinct columns of U are independent. For positive normalized columns, set

$$U_{v,z} = \log p(v \mid z)$$

Then

$$\text{softmax}(U_{\cdot,z})_v = \frac{\exp(\log p(v \mid z))}{\sum_{w \in \mathcal{V}} \exp(\log p(w \mid z))} = p(v \mid z).$$

For arbitrary probability columns and $0 < a < 1$, let

$$p^{(a)}(v \mid z) = (1-a)p(v \mid z) + \frac{a}{KM}.$$

Their finite log columns satisfy

$$\left|U_{v,z}^{(a)}\right| \leq \log(KM/a).$$

and converge simultaneously, since

$$\left|p^{(a)}(v \mid z) - p(v \mid z)\right| = a\left|\frac{1}{KM} - p(v \mid z)\right| \leq a.$$

The prescribed frozen columns are recovered in this probability limit; boundary zeros require the closure. □

D.1.2 Correct prefixes and information under occupancy

For a fixed closure policy and an unfinished tuple $z = (t, s, u, \perp)$, write $c(z) = \pi(C_{t+1,u} \mid z)$, and set $c(z) = 0$ at terminal tuples.

Lemma D.3 (Tuple-based seed sampling)**.** *Under the standing assumptions, fix any $\pi \in \mathcal{P}$. Its stopped student rollout can be realized using independent $R_0, \ldots, R_{K-1}$, each uniform on $[0,1)$ and independent of (X, B), so that the next token is a function only of its actual tuple and the current seed. At an unfinished tuple $z = (t, s, u, \perp)$, continuation occurs exactly when $R_m < c(z)$, and that continuation is $C_{t+1,u}$. At a terminal tuple every sampled token is an answer.*

Proof. Allocate $[0, c(z))$ to the allowed continuation and, for $j \in [M]$, allocate

$$\left[c(z) + \sum_{i<j} \pi(A_i \mid z),\ c(z) + \sum_{i\le j} \pi(A_i \mid z)\right).$$

to A_j. At terminal tuples set $c(z) = 0$. These intervals partition $[0, 1)$, with zero-mass intervals empty. Use fresh independent $R_m \sim \mathrm{Unif}[0, 1)$, independent of (X, B), at each visit. Conditional on the past, interval lengths equal the policy probabilities; multiplying them gives the prescribed stopped rollout law. □

Lemma D.4 (Correct surviving prefixes)**.** *Under the standing assumptions, fix any $\pi \in \mathcal{P}$. For each $0 \le t \le K-1$, let E_t be survival through t continuations without an answer and set*

$$\mathcal{F}_t = \sigma(S_0, f_1, \ldots, f_t, R_0, \ldots, R_{t-1}).$$

Then $E_t \in \mathcal{F}_t$, and on E_t the emitted prefix is exactly $C_{1,S_1}, \ldots, C_{t,S_t}$, the current stage/state is (t, S_t), and its decoded feature is $e_{(t,S_t,S_{t+1},\perp)}$. For $0 \le t < K-1$,

$$E_{t+1} = E_t \cap \left\{R_t < \pi(C_{t+1,S_{t+1}} \mid t, S_t, S_{t+1}, \perp)\right\}. \tag{D.1}$$

Stage t is visited exactly once on E_t and otherwise never; after terminal stage $K-1$ an answer is emitted. All conclusions hold pathwise for the seed construction.

Proof. Induct on t. Initially $E_0 = \Omega$, $S_0 \in \mathcal{F}_0$, and the decoded next state is $f_1(S_0) = S_1$. On E_t, the sole continuation is $C_{t+1,S_{t+1}}$; hence

$$E_{t+1} = E_t \cap \{R_t < \pi(C_{t+1,S_{t+1}} \mid t, S_t, S_{t+1}, \perp)\} \in \mathcal{F}_{t+1}.$$

Its literal state is $(t+1, S_{t+1})$. This proves the invariant and (D.1). Each continuation increments stage once; every alternative stops. Terminal support consists only of answers, so $1 \le \tau \le K$. The argument also holds for continuation probabilities zero or one. □

Lemma D.5 (Unread suffix counting)**.** *Under the standing assumptions, for every $0 \le t \le K-1$, define*

$$\mathcal{H}_t = \sigma(S_0, f_1, \ldots, f_{t+1}, R_0, \ldots, R_{t-1}).$$

For every $0 \le t \le K-2$, the suffix $(f_{t+2}, \ldots, f_K)$ is independent of $\mathcal{H}_t$, retaining its product-uniform distribution. For every event $H \in \mathcal{H}_t$ and $j \in [M]$,

$$\Pr(Y = j, H) = \frac{\Pr(H)}{M}. \tag{D.2}$$

In particular, if $\Pr(H) > 0$, then $\Pr(Y = j \mid H) = 1/M$.

Proof. Let a range over read-data atoms $(S_0, f_1, \ldots, f_{t+1})$, and let $D_a \subseteq [0, 1)^t$ be the seed section of H. For any suffix tuple g, primitive independence gives

$$\begin{aligned}\Pr(H, (f_{t+2}, \ldots, f_K) = g) &= \sum_a \Pr(a) \Pr(R_{<t} \in D_a)\, (M!)^{-(K-t-1)} \\ &= \Pr(H)\, (M!)^{-(K-t-1)}.\end{aligned} \tag{D.3}$$

For each fixed read atom, there are $(M!)^{K-t-2}(M-1)!$ suffixes mapping S_{t+1} to j, out of $(M!)^{K-t-1}$. Thus

$$\frac{(M!)^{K-t-2}(M-1)!}{(M!)^{K-t-1}} = \frac{1}{M}. \tag{D.4}$$

$$\Pr(Y = j, H) = \sum_a \Pr(a) \Pr(R_{<t} \in D_a) \frac{1}{M} = \frac{\Pr(H)}{M},$$

For $t = K-2$, the intermediate list is empty and the same count applies. No unread-suffix conclusion is needed at $t = K-1$. □

Proposition D.2 (Information under selected tuple occupancy)**.** *Under the standing assumptions, fix any $\pi \in \mathcal{P}$. For $z = (t, s, u, \perp)$, $0 \le t \le K-1$, let*

$$G_{t,s,u} = E_t \cap \{S_t = s, S_{t+1} = u\}.$$

Then

$$\sum_{m=0}^{\tau-1} \mathbf{1}\{z_m = (t,s,u,\perp)\} = \mathbf{1}_{G_{t,s,u}}, \qquad d_\pi(t,s,u,\perp) = \Pr(G_{t,s,u}). \tag{D.5}$$

For every tuple with positive occupancy, writing $\beta_0 = 1-\rho$, $\beta_1 = \rho$, its conditional joint law is

$$\Pr(B = b, Y = j \mid G_{t,s,u}) = \begin{cases} \beta_b/M, & t < K-1, \\ \beta_b \mathbf{1}\{j = u\}, & t = K-1, \end{cases} \quad b \in \{0,1\},\ j \in [M]. \tag{D.6}$$

Thus the bit retains its original law and is independent of Y under selected tuple occupancy, unfinished Y is uniform given survival and the actual readout feature, and terminal Y equals u. No conditional law is claimed at a zero-occupancy tuple.

Proof. The prefix invariant identifies the unique stage-t visit with $G_{t,s,u}$; taking expectations gives (D.5). Also $G_{t,s,u} \in \mathcal{H}_t$. For $t < K-1$, Lemma D.5 gives

$$\Pr(Y = j, G_{t,s,u}) = \Pr(G_{t,s,u})/M. \tag{D.7}$$

At $t = K-1$, $S_{t+1} = Y$, so

$$\Pr(Y = j, G_{K-1,s,u}) = \mathbf{1}\{j = u\} \Pr(G_{K-1,s,u}). \tag{D.8}$$

Since $B \perp (X, R_0, \dots, R_{K-1})$, for $G = G_{t,s,u}$,

$$\Pr(B = b, Y = j, G) = \Pr(B = b) \Pr(Y = j, G) = \beta_b \Pr(Y = j, G). \tag{D.9}$$

Divide by $\Pr(G) > 0$ to obtain (D.6). Exact one-hot decoding makes tuple conditioning and feature conditioning identical. Null-occupancy tuples are excluded. □

D.1.3 Stopped forward-KL projection

For a fixed π, write $\mathcal{O}_\pi = \{z : d_\pi(z) > 0\}$, $q_m = q_{0,B}(\cdot \mid X, h_m)$, and $p_z = p(\cdot \mid z)$. For a probability vector a, define $h(a) = -\sum_{v:a(v)>0} a(v) \log a(v)$. These are functions of the setup's occupancy, labels, and probability columns.

Lemma D.6 (Finite-alphabet Gibbs inequality)**.** *For any probability vectors a, b on the setting's finite vocabulary $\mathcal{V}$, with the setting's zero/infinite conventions,*

$$D_{\mathrm{KL}}(a\|b) \ge 0, \qquad D_{\mathrm{KL}}(a\|b) = 0 \iff a = b.$$

Also $0 \le h(a) \le \log(KM)$.

Proof. Write $S = \{v : a(v) > 0\}$. If $b(v) = 0 < a(v)$, the divergence is infinite. Otherwise $-\log x \ge 1 - x$ gives

$$\begin{aligned} D_{\mathrm{KL}}(a\|b) &= \sum_{v \in S} a(v)\big[-\log(b(v)/a(v))\big] \\ &\ge \sum_{v \in S} (a(v) - b(v)) = 1 - \sum_{v \in S} b(v) = \sum_{v \notin S} b(v) \ge 0. \end{aligned}$$

Equality requires $b(v)/a(v) = 1$ on S and zero mass outside S, hence $a = b$. Finally, with the uniform vector,

$$0 \le \sum_{v \in S} a(v) \log(a(v) KM) = \log(KM) - h(a).$$

Together with $-a(v) \log a(v) \ge 0$, this proves the entropy bounds. □

Lemma D.7 (Occupancy-selected teacher mean)**.** *Under the standing assumptions, fix any $\pi \in \mathcal{P}$. For $z \in \mathcal{O}_\pi$, define the full-vocabulary vector*

$$\bar{q}_z(v) = \frac{1}{d_\pi(z)} \mathbb{E}_{\mu_\pi, B} \sum_{m=0}^{\tau-1} \mathbf{1}\{z_m = z\} q_m(v). \tag{D.10}$$

It is a probability vector in the allowed column simplex, and for $z = (t, s, u, \perp)$,

$$\bar{q}_z = \begin{cases} (1-\rho)\delta_{C_{t+1,u}} + \dfrac{\rho}{M} \sum_{j=1}^{M} \delta_{A_j}, & 0 \le t < K-1, \\ \displaystyle\sum_{j=1}^{M} \big((1-\rho) Q_0(j \mid u) + \rho \mathbf{1}\{j = u\}\big) \delta_{A_j} = \sum_{j=1}^{M} Q_\rho(j \mid u) \delta_{A_j}, & t = K-1. \end{cases} \tag{D.11}$$

Proof. Nonnegativity is immediate, and label normalization gives

$$\sum_v \bar{q}_z(v) = \frac{\mathbb{E} \sum_{m<\tau} \mathbf{1}\{z_m = z\} \sum_v q_m(v)}{d_\pi(z)} = \frac{\mathbb{E} \sum_{m<\tau} \mathbf{1}\{z_m = z\}}{d_\pi(z)} = 1.$$

By the single-visit identity, $\bar{q}_z = \mathbb{E}[q_t \mid G_{t,s,u}]$. At unfinished tuples the label is $\delta_{C_{t+1,u}}$ for $B = 0$ and δ_{A_Y} for $B = 1$, with

$$\mathbb{E}[\mathbf{1}\{B = 1, Y = j\} \mid G_{t,s,u}] = \rho/M.$$

At terminal tuples $Y = u$, so the mean answer coordinate is

$$(1-\rho) Q_0(j \mid u) + \rho \mathbf{1}\{j = u\},$$

which proves (D.11). These normalized columns have the allowed supports. Only the selected joint law of (B, Y) is used, so the same bit may be reused across visits. □

Lemma D.8 (Stopped forward-KL decomposition)**.** *Under the four primitive assumptions and basic setup of Lemma D.7, Lemma D.2, Proposition D.1, and Proposition D.2, and Lemmas D.6 and D.7, fix any $\pi \in \mathcal{P}$. The constant*

$$C_\pi = \sum_{z \in \mathcal{O}_\pi} d_\pi(z) h(\bar{q}_z) - \mathbb{E}_{\mu_\pi, B} \sum_{m=0}^{\tau-1} h(q_m) \tag{D.12}$$

is finite and independent of p. For every $p \in \mathcal{P}$, as an equality of extended real numbers with a finite first term,

$$\mathcal{L}_\pi(p) = C_\pi + \sum_{z \in \mathcal{O}_\pi} d_\pi(z) D_{\mathrm{KL}}(\bar{q}_z \| p_z). \tag{D.13}$$

Null-occupancy tuples are omitted from this sum; thus no product $0 \cdot \infty$ is used.

Proof. The cap gives

$$\sum_{z \in \mathcal{Z}} d_\pi(z) = \mathbb{E} \sum_{m<\tau} \sum_{z \in \mathcal{Z}} \mathbf{1}\{z_m = z\} = \mathbb{E}\tau \le K. \tag{D.14}$$

Thus $|C_\pi| \le K \log(KM) < \infty$. Set

$$T_{z,v} = \sum_{m<\tau} \mathbf{1}\{z_m = z\} q_m(v), \qquad \mathbb{E} T_{z,v} = d_\pi(z) \bar{q}_z(v).$$

Zero means imply $T_{z,v} = 0$ almost surely; null tuples are likewise never visited. If $\bar{q}_z(v) > 0 = p_z(v)$ at an occupied tuple, both sides of (D.13) are $+\infty$. Otherwise all relevant logarithms are finite, and finite grouping yields

$$\begin{aligned} \mathcal{L}_\pi(p) &= -\mathbb{E} \sum_{m<\tau} h(q_m) + \mathbb{E} \sum_{m<\tau} \sum_{v : q_m(v) > 0} q_m(v)[-\log p_{z_m}(v)] \\ &= -\mathbb{E} \sum_{m<\tau} h(q_m) + \sum_{z \in \mathcal{O}_\pi} d_\pi(z) \sum_{v : \bar{q}_z(v) > 0} \bar{q}_z(v)[-\log p_z(v)]. \end{aligned} \tag{D.15}$$

Using

$$\sum_{v : \bar{q}_z(v) > 0} \bar{q}_z(v)[-\log p_z(v)] = h(\bar{q}_z) + D_{\mathrm{KL}}(\bar{q}_z \| p_z).$$

This gives (D.13). No product of zero occupancy with an infinite divergence is taken. □

Proposition D.3 (Occupied-column minimizers)**.** *Under the standing assumptions, every fixed $\pi \in \mathcal{P}$ satisfies*

$$\operatorname*{argmin}_{p\in\mathcal{P}} \mathcal{L}_\pi(p) = \{p \in \mathcal{P} : p(\cdot \mid z) = \bar{q}_z \text{ for every } z \in \mathcal{O}_\pi\}. \tag{D.16}$$

This set is nonempty, its loss is C_π, and $0 \le C_\pi \le K\log(KM)$. The unique minimizing probability vector at each occupied tuple is (D.11)*. At each unoccupied empty-reference tuple the loss imposes no condition beyond the original allowed simplex; nonempty-reference columns retain their primitive frozen values.*

Proof. Independent columns make $p^\star_z = \bar{q}_z$ feasible on occupied tuples, with arbitrary allowed null student columns and unchanged frozen columns. The decomposition gives

$$\mathcal{L}_\pi(p^\star) = C_\pi, \qquad \mathcal{L}_\pi(p) \ge C_\pi \quad (p \in \mathcal{P}).$$

At a minimizer,

$$0 = \sum_{z\in\mathcal{O}_\pi} d_\pi(z) D_{\mathrm{KL}}(\bar{q}_z \| p_z).$$

Since $d_\pi(z) > 0$ and KL is nonnegative, equality holds exactly when $p_z = \bar{q}_z$ on every occupied tuple. This proves (D.16); the entropy bounds imply $0 \le C_\pi \le K\log(KM)$. □

D.1.4 Coverage and self-consistent identification

Lemma D.9 (Uniform consecutive task pairs)**.** *Under the standing assumptions, for every $0 \le t \le K-1$ and $s, u \in [M]$,*

$$\Pr(S_t = s) = \frac{1}{M}, \qquad \Pr(S_t = s, S_{t+1} = u) = \frac{1}{M^2}. \tag{D.17}$$

Proof. Conditioned on $f_1, \dots, f_t$, their composition is a bijection, hence S_t remains uniform. Independence and uniformity of f_{t+1} then give

$$\Pr(S_t = s, f_{t+1}(s) = u) = \Pr(S_t = s)\frac{(M-1)!}{M!} = \frac{1}{M^2}.$$

Since $S_{t+1} = f_{t+1}(S_t)$, this proves the pair law for all $0 \le t < K$. □

Lemma D.10 (Survival from earlier continuation columns)**.** *Under the standing assumptions, fix any $\pi \in \mathcal{P}$ and $0 \le t \le K-1$. If*

$$\pi(C_{\ell+1,u} \mid \ell, s, u, \perp) = 1 - \rho \quad \text{for every } 0 \le \ell < t \text{ and } s, u \in [M], \tag{D.18}$$

then for every task input x and every $s, u \in [M]$,

$$\Pr_\pi(E_t \mid X = x) = (1-\rho)^t, \qquad d_\pi(t, s, u, \perp) = \frac{(1-\rho)^t}{M^2} > 0. \tag{D.19}$$

For $t = 0$, condition (D.18) *is empty.*

Proof. The survival recursion, with the prescribed earlier-stage continuation probabilities, gives by induction

$$E_t = \bigcap_{\ell=0}^{t-1} \{R_\ell < 1 - \rho\}. \tag{D.20}$$

Independent seeds imply $\Pr(E_t \mid X = x) = (1-\rho)^t$, including the empty product at $t = 0$. Therefore

$$\begin{aligned} d_\pi(t, s, u, \perp) &= \sum_x \Pr(X = x)\mathbf{1}\{S_t(x) = s, S_{t+1}(x) = u\} \Pr_\pi(E_t \mid X = x) \\ &= (1-\rho)^t \Pr(S_t = s, S_{t+1} = u) = \frac{(1-\rho)^t}{M^2}, \end{aligned} \tag{D.21}$$

This is positive because $\rho < 1$. □

Proposition D.4 (Successive fitted-stage identification)**.** *Under the standing assumptions, fix $\pi \in \mathcal{P}$ satisfying*

$$\pi \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathcal{L}_\pi(p). \tag{D.22}$$

For each $0 \le t \le K-1$, the following recurring assertion holds: for every $0 \le \ell \le t$ and $s, u \in [M]$,

$$\begin{aligned} \Pr_\pi(E_\ell \mid X = x) &= (1-\rho)^\ell \quad \text{for every } x, \\ d_\pi(\ell, s, u, \perp) &= \frac{(1-\rho)^\ell}{M^2} > 0, \\ \pi(\cdot \mid \ell, s, u, \perp) &= p_\rho(\cdot \mid \ell, s, u, \perp). \end{aligned} \tag{D.23}$$

Proof. At stage zero, $d_\pi(0, s, u, \perp) = M^{-2} > 0$. Self-consistency and Proposition D.3 therefore give $\pi_0 = p_{\rho,0}$, with continuation probability $1-\rho$. Assume the columns agree through stage $t < K-1$. Lemma D.10 yields

$$\Pr_\pi(E_{t+1} \mid X = x) = (1-\rho)^{t+1}, \qquad d_\pi(t+1, s, u, \perp) = \frac{(1-\rho)^{t+1}}{M^2} > 0$$

so the same minimizer characterization identifies all stage-$t+1$ columns. Induction reaches the terminal Q_ρ columns and proves (D.23). Coverage is obtained before projection at each stage. □

Proposition D.5 (Self-consistency of the declared candidate)**.** *Under the standing assumptions, the candidate p_ρ belongs to $\mathcal{P}$, satisfies*

$$p_\rho \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathcal{L}_{p_\rho}(p), \tag{D.24}$$

and has $d_{p_\rho}(t, s, u, \perp) = (1-\rho)^t / M^2$ for every student tuple.

Proof. Unfinished columns have nonnegative masses summing to $(1-\rho) + M(\rho/M) = 1$; terminal columns satisfy

$$Q_\rho(j \mid u) = (1-\rho) Q_0(j \mid u) + \rho \mathbf{1}\{j = u\} \ge 0, \qquad \sum_j Q_\rho(j \mid u) = (1-\rho) \cdot 1 + \rho = 1.$$

Together with the fixed reference columns, independent-column realization gives $p_\rho \in \mathcal{P}$. Its continuation probability is $1-\rho$, so its own occupancy is $(1-\rho)^t / M^2 > 0$. Its columns equal (D.11); Proposition D.3 proves self-consistency. □

Proposition D.6 (Zero-disclosure policy recovery)**.** *Under the standing assumptions, if $\rho = 0$, then the declared candidate and every self-consistently fitted π agree with p_0 on every student tuple. Their occupancies are exactly $1/M^2$, and every correct continuation through stage $K-1$ is emitted with probability one before an answer.*

Proof. At $\rho = 0$,

$$p_\rho(C_{t+1,u} \mid t, s, u, \perp) = 1 \ (t < K-1), \qquad Q_\rho = Q_0.$$

Frozen columns also equal the checkpoint, hence $p_{\rho=0} = p_0$. The fitted-policy characterization gives the same identity for every fitted policy, with occupancy M^{-2} and survival $\Pr(E_t) = 1$. □

D.1.5 Sampled execution and strict accuracy separation

Lemma D.11 (Terminal state and diagonal averaging)**.** *Under the standing assumptions, Y is uniform on $[M]$. For the candidate p_ρ, or for any fixed $\pi \in \mathcal{P}$ satisfying the self-consistent fit premise, $\Pr(E_{K-1}) = (1-\rho)^{K-1} > 0$ and*

$$\Pr(Y = u \mid E_{K-1}) = \frac{1}{M}, \qquad \mathbb{E}[Q_\rho(Y \mid Y) \mid E_{K-1}] = a_\rho. \tag{D.25}$$

The unconditional diagonal averages are $\mathbb{E}Q_0(Y \mid Y) = a_0$ and $\mathbb{E}Q_\rho(Y \mid Y) = a_\rho$.

Proof. The pair law gives

$$\Pr(Y = u) = \sum_{s=1}^{M} \Pr(S_{K-1} = s, S_K = u) = \sum_{s=1}^{M} \frac{1}{M^2} = \frac{1}{M}.$$

Task-independent survival gives

$$\Pr(E_{K-1}, Y = u) = \sum_{x} \Pr(X = x)\mathbf{1}\{Y(x) = u\}(1-\rho)^{K-1} = \frac{(1-\rho)^{K-1}}{M},$$
$$\Pr(E_{K-1}) = \sum_{x} \Pr(X = x)(1-\rho)^{K-1} = (1-\rho)^{K-1} > 0.$$

Thus $Y \mid E_{K-1}$ is uniform, and

$$\frac{1}{M}\sum_{u=1}^{M} Q_\rho(u \mid u) = (1-\rho)\frac{1}{M}\sum_{u=1}^{M} Q_0(u \mid u) + \rho = (1-\rho)a_0 + \rho = a_\rho. \tag{D.26}$$

□

Proposition D.7 (Sampled checkpoint teacher accuracy)**.** *Under the standing assumptions, the teacher's own sampled rollout satisfies*

$$A_T(p_0; \rho) = (1-\rho)a_0 + \rho = a_\rho. \tag{D.27}$$

Proof. For a fixed task, disclosure $b = 1$ produces A_Y immediately; $b = 0$ produces all correct continuations followed by $Q_0(\cdot \mid Y)$. Averaging the two kernels,

$$A_T(p_0; \rho) = \sum_{x} \Pr(X = x)\big[(1-\rho)Q_0(Y(x) \mid Y(x)) + \rho \cdot 1\big] = (1-\rho)a_0 + \rho,$$

by the uniform terminal law. □

Lemma D.12 (Correct forced prefixes)**.** *Under the standing assumptions, fix any task input x. For every $0 \le t \le K-1$, after t forced outputs p_ρ has itself generated exactly $C_{1,S_1(x)}, \ldots, C_{t,S_t(x)}$ with probability one, has emitted no answer, and has current stage/state $(t, S_t(x))$. The $t = 0$ assertion means the empty prefix and initial query. For every $0 \le t < K-1$, the next forced normalization has positive denominator $1-\rho$ and emits $C_{t+1,S_{t+1}(x)}$ with probability one, preserving the assertion at $t+1$. Thus after $K-1$ forced outputs the ordinary terminal tuple has retrieved state $Y(x)$.*

Proof. At each unfinished tuple, the total continuation mass is $1-\rho > 0$, entirely on $C_{t+1,u}$. Hence

$$\Pr_{\text{forced}}(C_{t+1,u} \mid t, s, u, \perp) = \frac{1-\rho}{1-\rho} = 1. \tag{D.28}$$

Starting from $(0, S_0)$, the implication

$$(t, S_t) \longmapsto C_{t+1,f_{t+1}(S_t)} = C_{t+1,S_{t+1}} \longmapsto (t+1, S_{t+1})$$

proves the generated-prefix claim by induction. □

Proposition D.8 (Self-generated forced execution)**.** *Under the standing assumptions, forced evaluation of p_ρ is well-defined at every visited position, generates $C_{1,S_1}, \ldots, C_{K-1,S_{K-1}}$ with probability one, and satisfies*

$$A_{\text{forced}}(p_\rho) = a_\rho. \tag{D.29}$$

Every fixed self-consistently fitted π has the same forced law and reward.

Proof. Lemma D.12 gives the correct chain through stage $K-1$. The unsuppressed final answer has law $Q_\rho(\cdot \mid Y)$, so

$$A_{\text{forced}}(p_\rho) = \mathbb{E}Q_\rho(Y \mid Y) = \frac{1}{M}\sum_{u} Q_\rho(u \mid u) = a_\rho.$$

Every fitted policy has identical student columns and thus identical forced kernels. □

Proposition D.9 (Completion and free first-answer reward)**.** *Under the standing assumptions,*

$$P_{\text{complete}}(p_\rho) = (1-\rho)^{K-1}, \qquad A_{\text{free}}(p_\rho) = \frac{1}{M} + \left(a_\rho - \frac{1}{M}\right)(1-\rho)^{K-1}. \tag{D.30}$$

Every fixed self-consistently fitted π has these same free sampled laws.

Proof. Let $D_t = E_t \setminus E_{t+1}$, $0 \le t < K-1$. These first-answer events and E_{K-1} partition the rollout space. In particular,

$$\Pr(D_t) = \Pr(E_t) - \Pr(E_{t+1}) = (1-\rho)^t - (1-\rho)^{t+1} = \rho(1-\rho)^t. \tag{D.31}$$

Fresh sampling seeds and the selected uniform law give

$$\begin{aligned}&\Pr(\text{current output} = A_Y \mid G_{t,s,u}) \\ &= \sum_{j=1}^{M} \Pr(Y=j \mid G_{t,s,u}) \Pr(\text{current output} = A_j \mid G_{t,s,u}, Y=j) = \sum_{j=1}^{M} \frac{1}{M}\frac{\rho}{M} = \frac{\rho}{M}.\end{aligned} \tag{D.32}$$

Summing over the stage-t tuple events gives the early-success contribution

$$\frac{\rho}{M}\Pr(E_t) = \frac{\rho}{M}(1-\rho)^t. \tag{D.33}$$

Terminal success contributes $(1-\rho)^{K-1}a_\rho$. Therefore

$$A_{\text{free}}(p_\rho) = \frac{1}{M}\sum_{t=0}^{K-2}\rho(1-\rho)^t + (1-\rho)^{K-1}a_\rho. \tag{D.34}$$

Since

$$\sum_{t=0}^{K-2}\rho(1-\rho)^t = \sum_{t=0}^{K-2}\left[(1-\rho)^t - (1-\rho)^{t+1}\right] = 1-(1-\rho)^{K-1}. \tag{D.35}$$

we obtain (D.30), including $\rho = 0$. Full student-column equality transfers these laws to every fitted policy. □

Proposition D.10 (Exact zero-disclosure reward recovery)**.** *Under the standing assumptions, if $\rho = 0$, then p_ρ and every self-consistently fitted π equal p_0 on every student tuple, and*

$$P_{\text{complete}}(p_{\rho=0}) = 1, \qquad A_{\text{free}}(p_{\rho=0}) = A_{\text{free}}(p_0) = a_0. \tag{D.36}$$

The corresponding fitted π has the same reward.

Proof. At zero disclosure the policy equals p_0 and completion is certain. The free-accuracy formula reduces to

$$\frac{1}{M} + \left(a_0 - \frac{1}{M}\right)\cdot 1 = a_0.$$

□

Proposition D.11 (Optional strict accuracy separation)**.** *Under the full stated assumption bases of Propositions D.7, D.8, and D.9, and additionally Assumption 6, fix K,M,Q_0 and vary $\rho \in [0,1)$. Then teacher and forced accuracies are strictly increasing, while fitted free accuracy is strictly decreasing. More explicitly, with $F(\rho) = A_{\text{free}}(p_\rho)$,*

$$\frac{d}{d\rho}A_T(p_0;\rho) = \frac{d}{d\rho}A_{\text{forced}}(p_\rho) = 1 - a_0 > 0, \tag{D.37}$$

$$F'(\rho) = (1-\rho)^{K-2}\left[(1-a_0) - (K-1)\left(a_0 - \frac{1}{M}\right) - K\rho(1-a_0)\right] < 0. \tag{D.38}$$

At $\rho = 0$ these derivatives have the right-derivative interpretation on the parameter domain.

Proof. The teacher and forced accuracies have derivative $1 - a_0 > 0$. For the free accuracy,

$$\begin{aligned} F(\rho) &= \frac{1}{M} + \left[a_0 - \frac{1}{M} + \rho(1-a_0)\right](1-\rho)^{K-1}. \\ F'(\rho) &= (1-a_0)(1-\rho)^{K-1} \\ &\quad - (K-1)\left[a_0 - \frac{1}{M} + \rho(1-a_0)\right](1-\rho)^{K-2} \\ &= (1-\rho)^{K-2}\left[(1-a_0) - (K-1)\left(a_0 - \frac{1}{M}\right) - K\rho(1-a_0)\right]. \end{aligned} \tag{D.39}$$

The threshold is equivalent to

$$Ka_0 > 1 + \frac{K-1}{M} \quad \Longleftrightarrow \quad (K-1)\left(a_0 - \frac{1}{M}\right) > (1-a_0). \tag{D.40}$$

so $F'(\rho) < 0$ for $0 \le \rho < 1$. Consequently, for $\rho_1 < \rho_2$,

$$F(\rho_2) - F(\rho_1) = \int_{\rho_1}^{\rho_2} F'(q)\, dq < 0.$$

□

D.1.6 Finite realization and actual reward transfer

In this subsection p denotes either p_ρ with reference $\perp$, or p_0 with the teacher reference r_B; p^ε denotes the corresponding declared smoothed policy. On every tuple, $p^\varepsilon = (1-\varepsilon)p + \varepsilon\nu$, since the candidate's nonempty-reference columns are frozen to the checkpoint.

Proposition D.12 (Finite unmasked smoothed columns)**.** *Under the standing assumptions, fix $0 < \varepsilon < 1$. Each declared smoothed policy $p_{0,\varepsilon}$ and $p_{\rho,\varepsilon}$ has strictly positive normalized columns and a finite unmasked realization with*

$$|\log p_{0,\varepsilon}(v \mid z)| \le \log(KM/\varepsilon), \qquad |\log p_{\rho,\varepsilon}(v \mid z)| \le \log(KM/\varepsilon) \tag{D.41}$$

for every vocabulary token v and every tuple z. Each smoothed column is at total variation distance at most ε from its ideal column. These columns are selected exactly at every actual prediction within the cap, including those after erroneous literal continuations. Teacher and initial student use identical finite columns, and the candidate's nonempty-reference columns remain those of that finite checkpoint.

Proof. For either prescribed ideal policy, smoothing gives

$$\begin{gathered} p^\varepsilon(v \mid z) = (1-\varepsilon)p(v \mid z) + \frac{\varepsilon}{KM}, \qquad \sum_{v \in \mathcal{V}} p^\varepsilon(v \mid z) = 1 - \varepsilon + \varepsilon = 1. \\ 0 < \frac{\varepsilon}{KM} \le p^\varepsilon(v \mid z) \le 1, \quad \log(\varepsilon/(KM)) \le \log p^\varepsilon(v \mid z) \le 0, \\ |\log p^\varepsilon(v \mid z)| \le \log(KM/\varepsilon). \end{gathered} \tag{D.42}$$

and

$$\frac{1}{2}\sum_{v \in \mathcal{V}} |p^\varepsilon(v \mid z) - p(v \mid z)| = \frac{\varepsilon}{2}\sum_{v \in \mathcal{V}} |\nu(v) - p(v \mid z)| \le \frac{\varepsilon}{2}\sum_{v \in \mathcal{V}} (\nu(v) + p(v \mid z)) = \varepsilon.$$

Assign $U_{v,z} = \log p^\varepsilon(v \mid z)$. Exact decoding selects these finite columns at every literal capped prefix, including malformed continuations. Teacher and initial student use identical checkpoint columns; candidate reference columns stay fixed. □

Lemma D.13 (Mixture sampling on the same task)**.** *Under the standing assumptions, each pair (p_ρ,$p_{\rho,\varepsilon}$) or (p_0,$p_{0,\varepsilon}$) admits a joint construction using the same X and B, whose marginals are the policies' actual capped sampled rollouts. The candidate pair uses reference $\perp$ in both copies; the teacher pair uses the same r_B in both. At each active finite-copy prediction the construction draws from its ideal column with probability $1 - \varepsilon$ and from ν with probability ε, independently of earlier random seeds.*

Proof. Enumerate the vocabulary and define $F_p(z, w) = v_i$ when

$$\sum_{j<i} p(v_j \mid z) \le w < \sum_{j\le i} p(v_j \mid z).$$

Take independent $J_m \sim \mathrm{Bernoulli}(\varepsilon)$, $W_m \sim \mathrm{Unif}[0,1)$, and $V_m \sim \nu$, independent of the shared (X, B). The ideal copy outputs $F_p(z, W_m)$; the finite copy outputs

$$\begin{cases} F_p(z^\varepsilon, W_m), & J_m = 0, \\ V_m, & J_m = 1. \end{cases} \tag{D.43}$$

using its own tuple z^ε. Its conditional law is

$$(1-\varepsilon)p(v \mid z^\varepsilon) + \varepsilon\nu(v) = p^\varepsilon(v \mid z^\varepsilon). \tag{D.44}$$

Each copy stops at its first answer or the cap and is held fixed thereafter. Fresh seeds give the correct marginal history laws. Use reference $\perp$ for the candidate pair and the common r_B for the teacher pair. □

Lemma D.14 (Matching histories before an override)**.** *Under the standing assumptions, let H_m and H_m^ε be the ideal and smoothed coupled histories after m potential emissions, including their actual emitted tokens and stopping status. Each history is held fixed once its copy has stopped at an answer or the cap. Define $G_m = \{J_1 = \cdots = J_m = 0\}$, with G_0 the full space. For every fixed task/reference kernel and every $0 \le m \le K$,*

$$\Pr(G_m) = (1-\varepsilon)^m, \qquad H_m = H_m^\varepsilon \text{ with the same stopping status on } G_m. \tag{D.45}$$

The assertion holds at the actual empty-prefix entry. Whenever it holds at $m < K$, the event G_{m+1} makes the next joint transition preserve it, including if both copies have already stopped.

Proof. Independence gives

$$\Pr(G_m) = \prod_{j=1}^{m} \Pr(J_j = 0) = (1-\varepsilon)^m,$$

with empty product one. Initially the histories coincide. If they coincide on G_m, then on G_{m+1} both active copies use the same tuple and $F_p(z, W_{m+1})$; stopped copies remain fixed. Thus the histories and stopping flags coincide by induction through $m = K$. □

Proposition D.13 (Transfer of actual capped rewards)**.** *Under the standing assumptions, for every $0 < \varepsilon < 1$, the coupled first-answer rewards satisfy*

$$\begin{aligned} |A_{\mathrm{free}}(p_{\rho,\varepsilon}) - A_{\mathrm{free}}(p_\rho)| &\le 1 - (1-\varepsilon)^K \le K\varepsilon, \\ |A_T(p_{0,\varepsilon}; \rho) - A_T(p_0; \rho)| &\le 1 - (1-\varepsilon)^K \le K\varepsilon. \end{aligned} \tag{D.46}$$

The finite rewards in these expressions include successes after erroneous continuations.

Proof. Let $R, R^\varepsilon \in \{0,1\}$ be the actual first-answer rewards, including post-error successes. Matching histories on G_K give

$$|R^\varepsilon - R| \le \mathbf{1}_{G_K^c}. \tag{D.47}$$

Hence

$$|\mathbb{E}R^\varepsilon - \mathbb{E}R| \le \mathbb{E}|R^\varepsilon - R| \le \Pr(G_K^c) = 1 - (1-\varepsilon)^K. \tag{D.48}$$

and

$$1 - (1-\varepsilon)^K = \sum_{m=0}^{K-1} \left[(1-\varepsilon)^m - (1-\varepsilon)^{m+1}\right] = \varepsilon \sum_{m=0}^{K-1} (1-\varepsilon)^m \le \varepsilon \sum_{m=0}^{K-1} 1 = K\varepsilon \tag{D.49}$$

The coupling marginals identify these expectations with the respective free or teacher accuracies. □

Proposition D.14 (Exact smoothed zero-disclosure identity)**.** *Under the standing assumptions, if* $\rho = 0$ *and* $0 < \varepsilon < 1$*, then*

$$p_{\rho=0,\varepsilon}(v \mid z) = p_{0,\varepsilon}(v \mid z) \quad \textit{for every } z \in \mathcal{Z},\ v \in \mathcal{V}. \tag{D.50}$$

Their finite log columns can be identical, and their resulting capped rollout laws coincide at any common reference, including on paths after finite-model errors.

Proof. Since $p_{\rho=0} = p_0$ on every tuple, identical smoothing gives

$$p_{\rho=0,\varepsilon}(v \mid z) = (1-\varepsilon)p_{\rho=0}(v \mid z) + \varepsilon\nu(v) = (1-\varepsilon)p_0(v \mid z) + \varepsilon\nu(v) = p_{0,\varepsilon}(v \mid z). \tag{D.51}$$

Identical columns imply identical capped rollout laws at every common reference. □

D.2 Theorem 2

Throughout this section, we work under the setting in Appendix B.4.

D.2.1 Exact finite neural selection

Lemma D.15 (Finite-head coordinate separation)**.** *Under the standing assumptions, let* h *be any prediction prefix before answer termination with next emitted position* $m \in \{1, \dots, K\}$*, let* $r \in \mathcal{R}$*, and let* (t, s) *be its literal current-token stage/state. Put* $u = f_{t+1}(s)$*. Then*

$$\frac{7}{8} \le a_u^{\mathrm{step}} \le 1, \qquad 0 \le a_i^{\mathrm{step}} \le \frac{1}{8} \quad (i \in [M],\ i \ne u),$$

$$\frac{7}{8} \le a_r^{\mathrm{ref}} \le 1, \qquad 0 \le a_{r'}^{\mathrm{ref}} \le \frac{1}{8} \quad (r' \in \mathcal{R},\ r' \ne r).$$

These conclusions hold with no correctness or ordering condition on earlier continuation tokens.

Proof. Before prediction $m \le K$, the visible context has

$$KM + 2 + (m-1) = KM + m + 1 \le KM + K + 1 \le N_{\mathrm{ctx}}$$

positions. Literal semantics give $(t, s) = (0, S_0)$ initially and (j, i) after any $C_{j,i}$. The unique table record $(t+1, s)$ has score 2β, while all competitors have score at most β. Hence

$$\alpha_*^{\mathrm{step}} = \frac{1}{1 + \sum_{q \ne *} e^{b_q - 2\beta}} \ge \frac{1}{1 + (N_{\mathrm{ctx}} - 1)e^{-\beta}}.$$

$$(N_{\mathrm{ctx}} - 1)e^{-\beta} \le \frac{N_{\mathrm{ctx}} - 1}{7(N_{\mathrm{ctx}} - 1)} = \frac{1}{7}, \qquad \alpha_*^{\mathrm{step}} \ge \frac{1}{1 + 1/7} = \frac{7}{8}.$$

Its value is e_u, so $a_u^{\mathrm{step}} \ge 7/8$; for $i \ne u$,

$$a_i^{\mathrm{step}} = \sum_{q \ne *} \alpha_q^{\mathrm{step}} \mathbf{1}_{\{q \text{ is a table record with value } i\}} \le \sum_{q \ne *} \alpha_q^{\mathrm{step}} = 1 - \alpha_*^{\mathrm{step}} \le \frac{1}{8}.$$

For the unique reference slot,

$$\alpha_*^{\mathrm{ref}} = \frac{1}{1 + \sum_{q \ne *} e^{-\beta}} \ge \frac{1}{1 + (N_{\mathrm{ctx}} - 1)e^{-\beta}} \ge \frac{7}{8}.$$

$$a^{\mathrm{ref}} = \alpha_*^{\mathrm{ref}} e_r.$$

These bounds hold at every literal prefix, including malformed ones. □

Lemma D.16 (Exact plateau decoding)**.** *Under the standing assumptions, at every capped prediction prefix with literal* (t, s) *and reference* $r \in \mathcal{R}$*, the decoded feature vector satisfies*

$$\phi_\zeta = \begin{cases} 1, & \zeta = (t, s, f_{t+1}(s), r), \\ 0, & \zeta \ne (t, s, f_{t+1}(s), r), \end{cases} \qquad \phi = e_{(t,s,f_{t+1}(s),r)}.$$

Proof. The disjoint residual blocks preserve the exact stage/state indicators. Thus

$$v_\zeta = \mathbf{1}_{\{t=t'\}} + \mathbf{1}_{\{s=s'\}} + a^{\text{step}}_{u'} + a^{\text{ref}}_{r'}.$$
$$v_{(t,s,u,r)} \geq 1 + 1 + \frac{7}{8} + \frac{7}{8} = \frac{15}{4}.$$

For $v \geq 15/4$,

$$2\{\text{ReLU}(v - 13/4) - \text{ReLU}(v - 15/4)\} = 2\{(v - 13/4) - (v - 15/4)\} = 1.$$

For a wrong stage/state, $v_\zeta \leq 3$; otherwise a wrong head coordinate gives $v_\zeta \leq 25/8 < 13/4$, so both ReLUs vanish. Therefore $\phi = e_{(t,s,u,r)}$. There are $KM^2(M+1)$ tuples and two hidden ReLUs per tuple, giving the stated width. □

Proposition D.15 (Literal-prefix readout realization)**.** *Under the standing assumptions, for every finite tied vector $\theta = (z_1, \ldots, z_d, w)$, every capped prediction prefix with literal (t, s), and every $r \in \mathcal{R}$, the actual network logits are exactly the prescribed feature-column coefficients at (t, s, u, r), where $u = f_{t+1}(s)$. This statement includes arbitrary incorrect-state and out-of-order continuation prefixes and both terminal and nonterminal literal stages.*

Proof. Only the decoded block has nonzero readout coefficients. Hence

$$\ell_v = \sum_\zeta W_{v,\zeta}\phi_\zeta = W_{v,(t,s,u,r)},$$

For $r = \perp, t < d$,

$$\ell_{C_{j,i}} = \begin{cases} z_{t+1} + \log M, & (j,i) = (t+1, u), \\ -L, & (j,i) \neq (t+1, u), \end{cases} \qquad \ell_{A_i} = 0.$$

For $r = \perp, t = d$,

$$\ell_{A_i} = \begin{cases} w + \log(M-1), & i = u, \\ 0, & i \neq u, \end{cases} \qquad \ell_{C_{j,i}} = -L \quad (j \in [d],\ i \in [M]).$$

For $r \in [M]$,

$$\ell_{A_i} = \begin{cases} r_A, & i = r, \\ 0, & i \neq r, \end{cases}$$

and, when $t < d$,

$$\ell_{C_{j,i}} = \begin{cases} r_C, & (j,i) = (t+1, u), \\ -L, & (j,i) \neq (t+1, u), \end{cases}$$

At $t = d$ all continuation logits equal $-L$. The selected tuple after $C_{j,i}$ is $(j, i, f_{j+1}(i), r)$, irrespective of emission position or correctness. Every row remains in the finite softmax. □

D.2.2 Comparison law, occupancy, and population iteration

Lemma D.17 (Comparison output laws)**.** *Under the standing assumptions, fix arbitrary finite $\theta = (z_1, \ldots, z_d, w)$, a capped prefix with literal (t, s), and put $u = f_{t+1}(s)$. Write $p_{\theta,\infty}$ for the setting's structural support comparison law. At empty reference and $t < d$, its only continuation is $C_{t+1,u}$, with*

$$p_{\theta,\infty}(C_{t+1,u} \mid h, \perp) = \sigma(z_{t+1}), \qquad p_{\theta,\infty}(A_i \mid h, \perp) = \frac{1 - \sigma(z_{t+1})}{M} \quad (i \in [M]).$$

At empty reference and $t = d$, it has no continuation and

$$p_{\theta,\infty}(A_u \mid h, \perp) = \sigma(w), \qquad p_{\theta,\infty}(A_i \mid h, \perp) = \frac{1 - \sigma(w)}{M - 1} \quad (i \neq u).$$

With any disclosed reference $r \in [M]$, at $t < d$ the only continuation again is $C_{t+1,u}$, with probability c_{ref}, and

$$p_{\theta,\infty}(A_r \mid h, r) = (1 - c_{\text{ref}})a_{\text{ref}}, \quad p_{\theta,\infty}(A_i \mid h, r) = \frac{(1 - c_{\text{ref}})(1 - a_{\text{ref}})}{M - 1} \quad (i \neq r).$$

At $t = d$, there are only answers with probabilities a_{ref} at r and $(1 - a_{\text{ref}})/(M - 1)$ elsewhere. All listed active probabilities are positive. The comparison teacher $q_{T,\infty}$ uses these laws at θ_0 and r_B.

Proof. Structural deletion leaves the empty-reference partitions

$$Me^{z_{t+1}} + M = M(1 + e^{z_{t+1}}).$$

$$(M-1)e^{w} + (M-1) = (M-1)(1 + e^{w}),$$

at nonterminal and terminal stages, respectively. For disclosed reference,

$$e^{r_A} + M - 1 = \frac{(M-1)a_{\mathrm{ref}}}{1 - a_{\mathrm{ref}}} + (M-1) = \frac{M-1}{1 - a_{\mathrm{ref}}},$$

$$e^{r_C} = \frac{c_{\mathrm{ref}}}{1 - c_{\mathrm{ref}}} \frac{M-1}{1 - a_{\mathrm{ref}}}.$$

Normalizing gives all stated laws. Interiority ensures positive active probabilities. Structural row identities determine deletion even if an active logit numerically equals $-L$. □

Proposition D.16 (True-path support and occupancy)**.** *Under the standing assumptions, fix a question X and a finite comparison vector θ. For each $1 \le m \le K$, let V_m mean that a free comparison student rollout reaches prediction m. On V_m its emitted tokens are exactly*

$$C_{1,S_1}, \ldots, C_{m-1,S_{m-1}},$$

its current tuple satisfies $(t, s) = (m-1, S_{m-1})$, and

$$\Pr(V_m \mid X) = \prod_{i=1}^{m-1} \sigma(z_i).$$

The product at $m = 1$ is one. Under position-forced comparison student evaluation, every prediction $m \le K$ is reached and has this true prefix. A free frozen comparison teacher has the same true-prefix property, with reach probability c_0^{m-1} for reference $\perp$, and c_{ref}^{m-1} for reference Y. Every one of these free or forced comparison rollouts emits an answer by position K.

Proof. Initially V_1 is certain. At a reached true prefix, the sole continuation is $C_{m,f_m(S_{m-1})} = C_{m,S_m}$, giving

$$\Pr(V_{m+1} \mid X) = \Pr(V_m \mid X)\sigma(z_m) = \prod_{i=1}^{m} \sigma(z_i).$$

Induction yields the true-prefix and occupancy claims. Terminal support consists only of answers. Forced renormalization chooses the same sole continuation with probability one. Replacing $\sigma(z_m)$ by c_0 or c_{ref} gives the two teacher visit laws. □

Lemma D.18 (Stopped tied-coordinate KL derivative)**.** *Under the standing assumptions, fix any capped prefix, privilege value B, and finite θ. In the comparison KL against the frozen teacher, at nonterminal literal stage $t < d$ the only nonzero possible tied derivative is*

$$\partial_{z_{t+1}} \mathrm{KL}(q_{T,\infty} \| p_{\theta,\infty}) = p_{\theta,\infty}(C_{t+1,u} \mid h, \perp) - q_{T,\infty}(C_{t+1,u} \mid h, B).$$

At $t = d$, the only nonzero possible tied derivative is

$$\partial_w \mathrm{KL}(q_{T,\infty} \| p_{\theta,\infty}) = p_{\theta,\infty}(A_u \mid h, \perp) - q_{T,\infty}(A_u \mid h, B).$$

Each displayed derivative has absolute value at most one.

Proof. On the common active support $\mathcal{S}$, only the designated logit depends on x, with $\partial_x \ell_v = \mathbf{1}\{v = v_*\}$. Let $Z = \sum_{v \in \mathcal{S}} e^{\ell_v}$. Holding the teacher and prefix fixed,

$$\mathrm{KL}(q \| p) = \sum_{v \in \mathcal{S}} q(v) \log q(v) - \sum_{v \in \mathcal{S}} q(v)\ell_v + \log Z,$$

$$\partial_x \mathrm{KL}(q \| p) = -q(v_*) + \frac{e^{\ell_{v_*}}}{Z} = p(v_*) - q(v_*).$$

All other tied derivatives vanish; the absolute derivative is at most one. Exact feature selection prevents multiplicity from unselected tied columns. □

Proposition D.17 (Exact comparison population field)**.** *Under Assumptions 7, 8, 9, and 10, Lemmas D.17 and D.18, and Proposition D.16, the comparison population semi-gradient at every finite $\theta \in \mathbb{R}^K$ is*

$$(g_\infty(\theta))_j = \left(\prod_{i<j} \sigma(z_i)\right)(\sigma(z_j) - c_*), \quad j \in [d], \qquad (g_\infty(\theta))_w = \left(\prod_{i=1}^{d} \sigma(z_i)\right)(\sigma(w) - a_*).$$

Here g_∞ is exactly the setting's expected token sum of stopped comparison KL derivatives, with one independent shared B, and each coordinate has absolute value at most one.

Proof. The teacher target means are

$$(1-\rho)c_0 + \rho c_{\mathrm{ref}} = c_*, \qquad (1-\rho)a_0 + \rho a_{\mathrm{ref}} = a_*.$$

The trajectory coordinate sums are, respectively,

$$\mathbf{1}_{V_j}\big[\sigma(z_j) - (1-B)c_0 - Bc_{\mathrm{ref}}\big],$$

$$\mathbf{1}_{V_K}\big[\sigma(w) - (1-B)a_0 - Ba_{\mathrm{ref}}\big].$$

Since B is independent of the student rollout, taking expectations multiplies each mean difference by its visit probability. Proposition D.16 therefore gives the claimed field. Its coordinates are bounded by one. The stopped rule differentiates neither occupancy nor teacher labels. □

Proposition D.18 (Comparison success probabilities)**.** *Under the standing assumptions, define $F_\infty(\theta), G_\infty(\theta)$ by the setting's free and position-forced success metrics applied to its comparison model at arbitrary finite θ. Then*

$$F_\infty(\theta) = u_0 + (\sigma(w) - u_0)\prod_{i=1}^{d} \sigma(z_i), \qquad G_\infty(\theta) = \sigma(w).$$

The frozen comparison teacher free accuracies, denoted $T_\infty(\perp)$ and $T_\infty(Y)$, satisfy

$$T_\infty(\perp) = A_0, \qquad T_\infty(Y) = a_{\mathrm{ref}}.$$

Proof. Conditional on any task, an early answer is uniform. Partitioning by the first-answer position gives

$$F_\infty(\theta) = u_0 \sum_{j=1}^{d} \left(\prod_{i<j} \sigma(z_i)\right)(1 - \sigma(z_j)) + \sigma(w)\prod_{i=1}^{d} \sigma(z_i).$$

The early-answer sum telescopes because its j-th factor is

$$\prod_{i<j} \sigma(z_i) - \prod_{i\le j} \sigma(z_i).$$

Thus $F_\infty = u_0 + (\sigma(w) - u_0)\prod_i \sigma(z_i)$. Forced generation reaches the true terminal prefix surely, giving $G_\infty = \sigma(w)$. Substitution of θ_0 gives A_0; for disclosed reference,

$$T_\infty(Y) = a_{\mathrm{ref}}\left[\sum_{j=1}^{d} c_{\mathrm{ref}}^{j-1}(1 - c_{\mathrm{ref}}) + c_{\mathrm{ref}}^{d}\right] = a_{\mathrm{ref}}[(1 - c_{\mathrm{ref}}^{d}) + c_{\mathrm{ref}}^{d}] = a_{\mathrm{ref}}.$$

□

Proposition D.19 (Finite comparison iteration)**.** *Under Assumptions 7, 8, 9, and 10, and Propositions D.17 and D.18, the initialized comparison population iteration has finite coordinates at every finite n. With exactly the setting's $c_j(n) = \sigma(\bar{z}_j(n))$ and $a(n) = \sigma(\bar{w}(n))$, its transitions are*

$$\bar{z}_j(n+1) = \bar{z}_j(n) + \eta\left(\prod_{i<j} c_i(n)\right)(c_* - c_j(n)),$$

$$\bar{w}(n+1) = \bar{w}(n) + \eta \left(\prod_{i=1}^{d} c_i(n) \right) (a_* - a(n)),$$

and its accuracies are

$$\bar{F}_n = u_0 + (a(n) - u_0) \prod_{i=1}^{d} c_i(n), \qquad \bar{G}_n = a(n).$$

These apply in particular to all recurrence indices $0 \le n < N$ and accuracy indices $0 \le n \le N$ for any finite N. At $\rho = 0$, the comparison sequence equals θ_0 at every finite index.

Proof. Interiority gives finite $\bar{\theta}_0$. Since $\|g_\infty\|_\infty \le 1$, every update is finite and the field/accuracy formulas apply inductively. The first update is

$$\bar{z}_j(1) = \mathrm{logit}(c_0) - \eta c_0^{j-1} \rho \Delta c, \qquad \bar{w}(1) = \mathrm{logit}(a_0) + \eta c_0^d \rho \Delta a.$$

At $\rho = 0$, $c_* = c_0, a_* = a_0$, so $g_\infty(\theta_0) = 0$ and $\bar{\theta}_n = \theta_0$ for all finite n. □

D.2.3 Comparison intervals, terminal clock, and optimization error

Lemma D.19 (Sigmoid steps cannot overshoot)**.** *Under the step-size condition $0 < \eta \le 1$ in Assumption 10, let $x \in \mathbb{R}$, $p \in (0,1)$, $v \in [0,1]$, $x_* = \mathrm{logit}(p)$, and $x^+ = x + \eta v(p - \sigma(x))$. Then x^+ lies in the closed segment from x to x_*, and*

$$|x^+ - x| \le \frac{\eta v}{4} |x_* - x| \le \frac{1}{4} |x_* - x|.$$

When $x = x_*$, the step is zero.

Proof. The global derivative bound is

$$\sigma'(y) = \frac{e^{-y}}{(1+e^{-y})^2} = \sigma(y)(1 - \sigma(y)) = \frac{1}{4} - (\sigma(y) - \tfrac{1}{2})^2 \in (0, 1/4].$$

For $x \ne x_*$, the mean value theorem gives $p - \sigma(x) = \sigma'(\xi)(x_* - x)$, hence

$$x^+ = (1 - \eta v \sigma'(\xi)) x + \eta v \sigma'(\xi) x_*.$$

Since $0 \le \eta v \sigma'(\xi) \le 1/4$, this is a convex combination. The case $x = x_*$ is stationary. □

Proposition D.20 (Comparison intervals and visit bounds)**.** *Under Assumptions 9 and 10, with $\rho = 0$ or the setting's strict $0 < \rho < 1$ regime, and Proposition D.19, Lemma D.19 implies the following recurring assertion at every finite n:*

$$\mathrm{logit}(c_*) \le \bar{z}_j(n) \le \mathrm{logit}(c_0), \qquad \mathrm{logit}(a_0) \le \bar{w}(n) \le \mathrm{logit}(a_*),$$
$$c_* \le c_j(n+1) \le c_j(n) \le c_0, \qquad a_0 \le a(n) \le a(n+1) \le a_*.$$

Furthermore

$$1 \ge \prod_{i<j} c_i(n) \ge c_*^{j-1} \ge c_*^{d-1} > 0, \qquad 1 \ge \prod_{i=1}^{d} c_i(n) \ge c_*^d > 0.$$

On the continuation logit interval $\sigma' \ge \lambda_c > 0$, and on the terminal logit interval $\sigma' \ge \lambda_a > 0$. In particular $0 < \gamma \le 1/4$. At $\rho = 0$, $\bar{\theta}_n = \theta_0$, $\bar{F}_n = A_0 = A_*$, and $\bar{G}_n = a_0 = a_*$ exactly.

Proof. Interiority gives

$$c_* = (1-\rho)c_0 + \rho c_{\mathrm{ref}} \in [c_{\mathrm{ref}}, c_0] \subset (0,1), \quad a_* = (1-\rho)a_0 + \rho a_{\mathrm{ref}} \in [a_0, a_{\mathrm{ref}}] \subset (u_0, 1).$$

Apply the no-overshoot lemma to each exact coordinate recurrence. Starting at the checkpoint endpoints, induction gives

$$\mathrm{logit}(c_*) \le \bar{z}_j(n+1) \le \bar{z}_j(n) \le \mathrm{logit}(c_0),$$
$$\mathrm{logit}(a_0) \le \bar{w}(n) \le \bar{w}(n+1) \le \mathrm{logit}(a_*).$$

Products yield the visit lower bounds. For $q = (1-t)l + tr$,

$$q(1-q) = (1-t)l(1-l) + tr(1-r) + t(1-t)(r-l)^2 \ge \min\{l(1-l), r(1-r)\}.$$

Applying this to the two probability intervals gives $\lambda_c, \lambda_a > 0$ and $0 < \gamma \le 1/4$. At $\rho = 0$ both intervals are singletons, proving the stated baseline identities. □

Proposition D.21 (Probability contraction and terminal clock control)**.** *Under Assumptions 9 and 10, the allowed baseline or strict ρ regime, Proposition D.19, and Proposition D.20, every finite $n \geq 0$ and $j \in [d]$ satisfy*

$$0 \leq c_j(n) - c_* \leq \rho\Delta c\, e^{-\eta\gamma n}, \qquad 0 \leq a_* - a(n) \leq \rho\Delta a\, e^{-\eta\gamma n},$$

and also

$$0 \leq a_* - a(n) \leq \rho\Delta a\, e^{-\lambda_a H_n}.$$

Proof. The mean value theorem on the invariant intervals gives

$$c_j(n+1) - c_* = \left[1 - \eta \left(\prod_{i<j} c_i(n)\right) \sigma'(\xi_{j,n})\right] (c_j(n) - c_*).$$

The bracket is nonnegative and at most

$$1 - \eta\lambda_c c_*^{j-1} \leq 1 - \eta\lambda_c c_*^{d-1} \leq 1 - \eta\gamma.$$

Likewise,

$$a_* - a(n+1) = \left[1 - \eta \left(\prod_{i=1}^{d} c_i(n)\right) \sigma'(\xi_{a,n})\right] (a_* - a(n)).$$

whose bracket is nonnegative and bounded by both

$$1 - \eta\lambda_a c_*^d \leq 1 - \eta\gamma \quad \text{and by} \quad 1 - \eta\lambda_a \prod_{i=1}^{d} c_i(n).$$

Using the initial errors $\rho\Delta c, \rho\Delta a$, iteration yields

$$c_j(n) - c_* \leq \rho\Delta c(1-\eta\gamma)^n, \quad a_* - a(n) \leq \rho\Delta a(1-\eta\gamma)^n,$$

$$a_* - a(n) \leq \rho\Delta a \prod_{k=0}^{n-1} \left(1 - \eta\lambda_a \prod_{i=1}^{d} c_i(k)\right).$$

Apply $1 - t \leq e^{-t}$ to obtain the exponential rates and the clock bound, since

$$\exp\left(-\lambda_a \eta \sum_{k=0}^{n-1} \prod_{i=1}^{d} c_i(k)\right) = e^{-\lambda_a H_n}.$$

Empty products and zero initial errors cover $n = 0$ and $\rho = 0$. □

Lemma D.20 (Two lower bounds for terminal visits)**.** *Under Assumptions 9 and 10, the allowed ρ regime, Proposition D.19, and Proposition D.20, every finite $n \geq 0$ and $j \in [d]$ satisfy*

$$\bar{z}_j(n) \geq \operatorname{logit}(c_0) - \eta n \rho \Delta c.$$

Consequently, for every finite integer $N \geq 0$,

$$H_N \geq \eta \sum_{n=0}^{N-1} \sigma(\operatorname{logit}(c_0) - \eta n \rho \Delta c)^d, \qquad H_N \geq \eta N c_*^d.$$

At $\rho = 0$, both lower bounds equal $H_N = \eta N c_0^d$.

Proof. The invariant intervals give

$$0 \leq \bar{z}_j(k) - \bar{z}_j(k+1) = \eta \left(\prod_{i<j} c_i(k)\right) (c_j(k) - c_*) \leq \eta(c_0 - c_*) = \eta\rho\Delta c.$$

Summation and monotonicity of σ yield

$$\bar{z}_j(n) \geq \operatorname{logit}(c_0) - \eta n \rho \Delta c, \qquad \prod_{j=1}^{d} c_j(n) \geq \sigma(\operatorname{logit}(c_0) - \eta n \rho \Delta c)^d.$$

Summing over $n < N$ proves the first clock bound. The second follows from $\prod_j c_j(n) \geq c_*^d$. At $\rho = 0$, both are equalities. □

Proposition D.22 (Comparison free-accuracy residual)**.** *Under the standing assumptions, for every finite $n \geq 0$,*

$$|\bar{F}_n - A_*| \leq (\rho\Delta a + d\rho\Delta c)e^{-\eta\gamma n} = B_{\mathrm{err}} e^{-\eta\gamma n}.$$

Proof. Subtract the exact accuracy formulas:

$$\bar{F}_n - A_* = (a(n) - a_*) \prod_{i=1}^{d} c_i(n) + (a_* - u_0) \left(\prod_{i=1}^{d} c_i(n) - c_*^d \right).$$

The product telescopes as

$$\prod_{i=1}^{d} c_i(n) - c_*^d = \sum_{j=1}^{d} (c_j(n) - c_*) \left(\prod_{i<j} c_i(n) \right) c_*^{d-j}.$$

Thus the contraction bounds give

$$0 \leq \prod_{i=1}^{d} c_i(n) - c_*^d \leq \sum_{j=1}^{d} (c_j(n) - c_*) \leq d\rho\Delta c\, e^{-\eta\gamma n}.$$

and, since $\prod_i c_i(n) \leq 1$ and $0 < a_* - u_0 < 1$,

$$|\bar{F}_n - A_*| \leq |a(n) - a_*| \cdot 1 + 1 \cdot \left| \prod c_i(n) - c_*^d \right| \leq (\rho\Delta a + d\rho\Delta c)e^{-\eta\gamma n}.$$

□

D.2.4 Finite-network gradients and tied-coordinate perturbation

Write g_L for the exact expected stopped token-sum gradient of the finite network; g_∞ is its structural comparison counterpart from Proposition D.17. Throughout this subsection and the next, use the explicit quantities

$$s_L = KMe^{-L}, \qquad J = 1 + 4\eta K^3, \qquad b_n = 4\eta K^3 M n J^n e^{-L}, \qquad \delta_n = \|\theta_n - \bar{\theta}_n\|_\infty.$$

These are defined expressions; the bounds involving them are proved below. In particular $\delta_N = \delta_{\theta,N}$ and $b_0 = 0$.

Lemma D.21 (Uniform structural-tail bound)**.** *Under the standing assumptions, for every finite θ, every capped prediction prefix h, and every $r \in \mathcal{R}$,*

$$\mathrm{TV}\big(p_{\theta,L}(\cdot \mid h, r), p_{\theta,\infty}(\cdot \mid h, r)\big) \leq s_L.$$

Here $\mathrm{TV}(p, q) = \frac{1}{2}\sum_v |p(v) - q(v)|$, with the comparison law extended by zero on deleted rows. In particular, uniformly in B,

$$\mathrm{TV}\big(q_{T,L}(\cdot \mid h, B), q_{T,\infty}(\cdot \mid h, B)\big) \leq s_L.$$

Proof. Let Z and T be the total active and suppressed weights. There are at most KM suppressed rows, so $T \leq KMe^{-L} = s_L$. At least $M - 1$ answer weights equal one, hence $Z \geq 1$, uniformly over parameters and literal prefixes. Active probabilities are rescaled by $Z/(Z + T)$, giving

$$\mathrm{TV}(p_{\theta,L}, p_{\theta,\infty}) = \frac{T}{Z + T} \leq \frac{T}{Z} \leq s_L.$$

Singleton differences are bounded by TV. The same argument applies to both teacher references. □

Lemma D.22 (Capped stopped-gradient bounds)**.** *Under the standing assumptions, let $D_L(h, B; \theta)$ be the actual fixed-prefix stopped KL gradient and $D_\infty(h, B; \theta)$ its structural comparison version. At every finite vector, capped prefix and bit,*

$$\|D_L\|_\infty, \|D_\infty\|_\infty \leq 1, \qquad \|D_L - D_\infty\|_\infty \leq 2s_L.$$

For each actual or comparison capped trajectory its corresponding token-sum gradient has norm at most K, and consequently $\|g_L(\theta)\|_\infty, \|g_\infty(\theta)\|_\infty \leq K$.

Proof. At each fixed prefix, only the designated logit depends on its selected tied coordinate, with derivative one. Differentiating

$$\sum_v q(v)\log q(v) - \sum_v q(v)\ell_v + \log\left(\sum_v e^{\ell_v}\right).$$

gives $D_L = (p_L(v_*) - q_L(v_*))e_x$. Thus

$$\|D_L\|_\infty, \|D_\infty\|_\infty \le 1, \qquad \|D_L - D_\infty\|_\infty \le 2s_L.$$

The last bound includes both student and teacher tails. Summing at most K visits gives trajectory and population norm bounds K, including repeated coordinates on malformed paths. □

Proposition D.23 (Actual finite-iterate box)**.** *Under the standing assumptions, the actual population iteration is finite at every finite index and satisfies*

$$\|\theta_n\|_\infty \le \|\theta_0\|_\infty + \eta K n = \max\{|\operatorname{logit}(c_0)|, |\operatorname{logit}(a_0)|\} + \eta K n.$$

Consequently $\|\theta_n\|_\infty \le R_N$ *for each integer* $0 \le n \le N$.

Proof. Starting from finite θ_0, the capped-gradient bound gives inductively

$$\|\theta_{n+1}\|_\infty \le \|\theta_n\|_\infty + \eta\|g_L(\theta_n)\|_\infty \le \|\theta_0\|_\infty + \eta K n + \eta K.$$

Hence $\|\theta_n\|_\infty \le \|\theta_0\|_\infty + \eta K n \le R_N$ for $n \le N$. □

Proposition D.24 (Whole-trajectory gradient comparison)**.** *Under Assumptions 8, 9, and 10, including its prescribed independent once-shared bit, and Lemmas D.21 and D.22, every finite* θ *satisfies*

$$\|g_L(\theta) - g_\infty(\theta)\|_\infty \le 2Ks_L + 2K^2 s_L \le 4K^3 M e^{-L}.$$

Proof. At common active prefixes, couple the two categorical laws using shared mass $\mu(v) = \min\{p(v), q(v)\}$, with $\alpha = \sum_v \mu(v)$. Sample the common part jointly and the residual parts separately. The mismatch probability is

$$1 - \alpha = \frac{1}{2}\sum_v |p(v) - q(v)| \le s_L$$

Pad stopped strings, and after mismatch run each copy with its own literal kernels. The first-discrepancy event $\mathcal{D}$ satisfies

$$\Pr(\mathcal{D} \mid X) \le \sum_{m=1}^{K} s_L = Ks_L.$$

Use one common B, independent of the coupling and task, for all labels. Define

$$S_L = \sum_{h \in \mathcal{H}(\tau_L)} D_L(h, B; \theta), \qquad S_\infty = \sum_{h \in \mathcal{H}(\tau_\infty)} D_\infty(h, B; \theta).$$

On $\mathcal{D}^c$, $\|S_L - S_\infty\|_\infty \le 2Ks_L$; on $\mathcal{D}$, the cap gives $\|S_L - S_\infty\|_\infty \le 2K$. Therefore

$$\begin{aligned}\|g_L(\theta) - g_\infty(\theta)\|_\infty &\le \mathbb{E}\|S_L - S_\infty\|_\infty \\ &\le 2Ks_L \Pr(\mathcal{D}^c) + 2K\Pr(\mathcal{D}) \\ &\le 2Ks_L + 2K^2 s_L.\end{aligned}$$

$$2Ks_L + 2K^2 s_L = 2K^2(1+K)Me^{-L} \le 4K^3 M e^{-L}.$$

The latter event includes all post-mismatch gradients; no malformed suffix is omitted. □

Lemma D.23 (Global comparison-field regularity)**.** *Under the standing assumptions, for every pair* $\theta, \theta' \in \mathbb{R}^K$,

$$\|g_\infty(\theta) - g_\infty(\theta')\|_\infty \le \frac{K}{4}\|\theta - \theta'\|_\infty \le 4K^3\|\theta - \theta'\|_\infty.$$

Proof. The field is globally continuously differentiable. Its nonzero derivatives are

$$\partial_{z_j}(g_\infty)_j = \left(\prod_{i<j}\sigma(z_i)\right)\sigma'(z_j), \qquad |\partial_{z_j}(g_\infty)_j| \le \frac{1}{4}.$$

$$\partial_{z_i}(g_\infty)_j = \sigma'(z_i)\left(\prod_{\substack{\ell<j\\ \ell\neq i}}\sigma(z_\ell)\right)(\sigma(z_j)-c_*), \qquad |\partial_{z_i}(g_\infty)_j| \le \frac{1}{4}.$$

for $i<j$, and

$$\partial_w(g_\infty)_w = \left(\prod_{i=1}^{d}\sigma(z_i)\right)\sigma'(w), \qquad |\partial_w(g_\infty)_w| \le \frac{1}{4},$$

$$\partial_{z_i}(g_\infty)_w = \sigma'(z_i)\left(\prod_{\substack{\ell\le d\\ \ell\neq i}}\sigma(z_\ell)\right)(\sigma(w)-a_*), \qquad |\partial_{z_i}(g_\infty)_w| \le \frac{1}{4}.$$

for the terminal row. Each row sum is at most $K/4$. Integrating along the segment between θ' and θ,

$$|(g_\infty(\theta)-g_\infty(\theta'))_k| \le \int_0^1 \sum_{\ell=1}^{K} |\partial_\ell(g_\infty)_k(\theta'+t(\theta-\theta'))|\,|\theta_\ell-\theta'_\ell|\,dt \le \frac{K}{4}\|\theta-\theta'\|_\infty.$$

Taking the maximum and using $K/4 \le 4K^3$ proves the claim. □

Proposition D.25 (Accumulated tied-coordinate comparison)**.** *Under the standing assumptions, for every fixed finite integer $N \ge 0$ and every $0 \le n \le N$, both iterates are finite and*

$$\delta_n \le 4\eta K^3 M e^{-L}\sum_{r=0}^{n-1} J^r \le b_n.$$

The sum is zero at $n=0$. At each $0 \le n < N$,

$$\delta_{n+1} \le J\delta_n + 4\eta K^3 M e^{-L}.$$

Proof. Subtract the finite and comparison updates and use the global bounds:

$$\begin{aligned}\delta_{n+1} &\le \delta_n + \eta\|g_L(\theta_n) - g_\infty(\bar\theta_n)\|_\infty\\ &\le \delta_n + \eta\|g_L(\theta_n)-g_\infty(\theta_n)\|_\infty + \eta\|g_\infty(\theta_n)-g_\infty(\bar\theta_n)\|_\infty\\ &\le (1+4\eta K^3)\delta_n + 4\eta K^3 M e^{-L}.\end{aligned}$$

Since $\delta_0 = 0$, induction with $J = 1+4\eta K^3$ gives

$$\delta_n \le 4\eta K^3 M e^{-L}\sum_{r=0}^{n-1} J^r.$$

For $n \ge 1$,

$$4\eta K^3 M e^{-L}\sum_{r=0}^{n-1} J^r \le 4\eta K^3 M n J^n e^{-L} = b_n.$$

The empty sum gives the same claim at $n=0$. □

D.2.5 TRANSFER TO FREE AND POSITION-FORCED SUCCESS PROBABILITIES

Lemma D.24 (Free success under structural deletion)**.** *Under the standing assumptions, for every finite θ,*

$$|F_L(\theta) - F_\infty(\theta)| \le K s_L = K^2 M e^{-L}.$$

For the frozen teacher, at the identical finite checkpoint and in each reference mode,

$$|T_L(\perp) - T_\infty(\perp)| = |T_L(\perp) - A_0| \le K^2 M e^{-L},$$
$$|T_L(Y) - T_\infty(Y)| = |T_L(Y) - a_{\mathrm{ref}}| \le K^2 M e^{-L}.$$

Proof. Use the common-mass coupling from Proposition D.24 at the same parameter and reference. Its first-discrepancy event satisfies

$$\Pr(\mathcal{D} \mid X) \le K s_L.$$

Let I_X be the actual first-answer success indicator, including post-error recovery. Then

$$|I_X(\tau_L) - I_X(\tau_\infty)| \le \mathbf{1}_{\mathcal{D}}, \qquad |\mathbb{E} I_X(\tau_L) - \mathbb{E} I_X(\tau_\infty)| \le K s_L.$$

Average over tasks. Taking $\theta = \theta_0$ with references $\perp$ and Y gives the teacher bounds against A_0 and a_{ref}, respectively. □

Proposition D.26 (Forced-prefix agreement and denominator)**.** *Under the standing assumptions, fix finite $N \ge 0$, a finite θ with $\|\theta\|_\infty \le R_N$, and a question X. Couple actual and comparison position-forced empty-reference rollouts. For $1 \le m \le K$, let $\mathcal{A}_m$ be agreement through the first $m-1$ outputs. There is a coupling for which on $\mathcal{A}_m$ both prefixes are $C_{1,S_1}, \dots, C_{m-1,S_{m-1}}$, and for every $m \le d$ the actual continuation-only partition at this prefix is at least $M e^{-R_N} > 0$. The probability of first disagreement at position m, conditional on X, is at most $K e^{R_N - L}$. Consequently*

$$\Pr(\mathcal{A}_K^c \mid X) \le dK e^{R_N - L}.$$

Proof. Before the first mismatch, induction gives the true prefix $(m-1, S_{m-1})$. At $m \le d$, cancellation of the full-softmax denominator gives continuation partition and correct-token probability

$$M e^{z_m} + (dM - 1) e^{-L}, \qquad \frac{M e^{z_m}}{M e^{z_m} + (dM-1) e^{-L}}.$$

Since $z_m \ge -R_N$,

$$M e^{z_m} + (dM-1) e^{-L} \ge M e^{z_m} \ge M e^{-R_N} > 0.$$

The conditional mismatch probability is at most

$$\frac{(dM-1) e^{-L}}{M e^{z_m} + (dM-1) e^{-L}} \le \frac{KM e^{-L}}{M e^{-R_N}} = K e^{R_N - L}.$$

Summing disjoint first-mismatch events gives $\Pr(\mathcal{A}_K^c \mid X) \le dK e^{R_N - L}$. After mismatch the actual process uses its own positive finite continuation weights; no comparison normalization is imposed on a malformed terminal prefix. □

Proposition D.27 (Position-forced success transfer)**.** *Under the standing assumptions, for every finite θ with $\|\theta\|_\infty \le R_N$,*

$$|G_L(\theta) - G_\infty(\theta)| \le dK e^{R_N - L} + s_L \le K^2 e^{R_N - L} + s_L.$$

Proof. On $\mathcal{A}_K$, both copies reach the true terminal prefix. Couple their ordinary final-token laws with mismatch probability at most s_L. Hence

$$\Pr(\mathcal{D} \mid X) \le \Pr(\mathcal{A}_K^c \mid X) + s_L \Pr(\mathcal{A}_K \mid X) \le dK e^{R_N - L} + s_L.$$

The actual success indicators differ by at most $\mathbf{1}_{\mathcal{D}}$, including recovery after wrong continuations. Expectations give the bound; use $d \le K$ for its second form. □

Lemma D.25 (Comparison accuracy regularity)**.** *Under Proposition D.18 in the setting with $M, K \ge 2$, for every finite $\theta, \theta' \in \mathbb{R}^K$,*

$$|F_\infty(\theta) - F_\infty(\theta')| \le \frac{K}{4} \|\theta - \theta'\|_\infty, \qquad |G_\infty(\theta) - G_\infty(\theta')| \le \frac{1}{4} \|\theta - \theta'\|_\infty.$$

Proof. Differentiation gives

$$\partial_w F_\infty = \sigma'(w)\prod_{j=1}^{d}\sigma(z_j), \qquad \partial_{z_i} F_\infty = (\sigma(w)-u_0)\sigma'(z_i)\prod_{\substack{1\le j\le d\\ j\neq i}}\sigma(z_j).$$

Since $\sigma' \le 1/4$,

$$\sum_{i=1}^{d}|\partial_{z_i}F_\infty| + |\partial_w F_\infty| \le \frac{d+1}{4} = \frac{K}{4}.$$

Also $\|\nabla G_\infty\|_1 \le 1/4$. For either function,

$$|H(\theta)-H(\theta')| \le \int_0^1 \sum_{i=1}^{K}|\partial_i H(\theta'+t(\theta-\theta'))|\,\|\theta-\theta'\|_\infty\,dt$$

which proves the two Lipschitz bounds. □

Proposition D.28 (Explicit endpoint accuracy budget)**.** *Under the standing assumptions, for each finite integer $N \ge 0$ and $n \in \{0, N\}$,*

$$\delta_{F,n} \le K^2 M e^{-L} + \frac{K}{4}b_n, \qquad \delta_{G,n} \le K^2 e^{R_N - L} + KMe^{-L} + \frac{1}{4}b_n.$$

All four endpoint errors and $\delta_{\theta,N}$ are at most the exact setting expression

$$E_N(L) = 20K^4 M(N+1)e^{R_N}J^N e^{-L}, \qquad J = 1+4\eta K^3.$$

Both frozen-teacher same-checkpoint free errors are at most K^2Me^{-L}, and consequently at most $E_N(L)$.

Proof. Insert the same-parameter comparison accuracies and apply the coupling and Lipschitz bounds:

$$\begin{aligned}\delta_{F,n} &\le |F_L(\theta_n)-F_\infty(\theta_n)| + |F_\infty(\theta_n)-F_\infty(\bar\theta_n)|\\ &\le K^2Me^{-L} + \frac{K}{4}\|\theta_n-\bar\theta_n\|_\infty \le K^2Me^{-L}+\frac{K}{4}b_n.\end{aligned}$$

$$\begin{aligned}\delta_{G,n} &\le |G_L(\theta_n)-G_\infty(\theta_n)| + |G_\infty(\theta_n)-G_\infty(\bar\theta_n)|\\ &\le K^2e^{R_N-L} + s_L + \frac{1}{4}\|\theta_n-\bar\theta_n\|_\infty \le K^2e^{R_N-L}+KMe^{-L}+\frac{1}{4}b_n.\end{aligned}$$

Put

$$C_N = K^4M(N+1)e^{R_N}J^Ne^{-L} > 0.$$

For $n \in \{0, N\}$, $b_n \le b_N$, and direct division gives

$$\frac{b_N}{C_N} \le 2, \qquad \frac{(K/4)b_N}{C_N} \le 1, \qquad \frac{b_N/4}{C_N} \le \frac{1}{2},$$

$$\frac{K^2Me^{-L}}{C_N} \le \frac{1}{4}, \qquad \frac{K^2e^{R_N-L}}{C_N} \le \frac{1}{8}, \qquad \frac{KMe^{-L}}{C_N} \le \frac{1}{8}.$$

Here $K, M \ge 2$, $R_N \ge 0$, $J \ge 1$, and $\eta \le 1$. Consequently

$$\delta_{\theta,N} \le b_N \le 2C_N \le 20C_N, \qquad \delta_{F,n} \le (\tfrac{1}{4}+1)C_N = \tfrac{5}{4}C_N \le 20C_N,$$

$$\delta_{G,n} \le (\tfrac{1}{8}+\tfrac{1}{8}+\tfrac{1}{2})C_N = \tfrac{3}{4}C_N \le 20C_N.$$

Since $E_N(L) = 20C_N$, all claimed budgets follow. Each teacher error is at most $K^2Me^{-L} \le C_N/4$. □

D.2.6 STRICT SEPARATION AND FINITE BUDGET SPECIALIZATION

Lemma D.26 (Primitive positivity of the execution gap)**.** *Under Assumptions 9 and 11, with the original $M, K \geq 2$, the setting expressions satisfy*

$$D = A_0 - A_* > 0, \qquad 0 < \gamma \leq 1/4, \qquad \rho\Delta a > 0, \qquad B_{\mathrm{err}} > 0,$$

and all these quantities and the checkpoint logits are finite. This conclusion uses no numerical budget.

Proof. Interiority and $0 < \rho < 1$ give

$$0 < c_{\mathrm{ref}} < c_* = (1-\rho)c_0 + \rho c_{\mathrm{ref}} < c_0 < 1, \qquad u_0 < a_0 < a_* = (1-\rho)a_0 + \rho a_{\mathrm{ref}} < a_{\mathrm{ref}} < 1.$$

Set $h(r) = u_0 + (a_0 + r\Delta a - u_0)(c_0 - r\Delta c)^d$. For $0 \leq r \leq \rho$,

$$\begin{aligned} h'(r) &= \Delta a(c_0 - r\Delta c)^d - d\Delta c(a_0 + r\Delta a - u_0)(c_0 - r\Delta c)^{d-1} \\ &= (c_0 - r\Delta c)^{d-1}\left[\Delta a\, c_0 - d\Delta c(a_0 - u_0) - (d+1)r\Delta a\Delta c\right] < 0. \end{aligned}$$

by the primitive strict inequality. Thus $D = h(0) - h(\rho) > 0$. Interiority also gives $\lambda_c, \lambda_a > 0$, hence $0 < \gamma \leq 1/4$, while $B_{\mathrm{err}} = \rho(\Delta a + d\Delta c) > 0$. □

Proposition D.29 (Existence of finite numerical budgets)**.** *Under the standing assumptions, there exists a finite integer N obeying the exact N threshold in Assumption 12. For every finite integer obeying that threshold, the finite choice*

$$L \geq \max\left\{1,\ R_N + N\log(1 + 4\eta K^3) + \log\frac{160K^4M(N+1)}{\min\{D, \rho\Delta a\}}\right\}$$

satisfies $L > 0$ and the exact $E_N(L)$ budget. Thus both finite numerical budgets are jointly attainable.

Proof. All quantities in the stated choices of N, L are finite and their logarithm arguments are positive. Taking the ceiling gives a finite integer N; the stated finite L satisfies

$$e^{-L} \leq \frac{\min\{D, \rho\Delta a\}}{160K^4M(N+1)e^{R_N}(1 + 4\eta K^3)^N}.$$

Therefore

$$E_N(L) = 20K^4M(N+1)e^{R_N}(1 + 4\eta K^3)^N e^{-L} \leq \frac{20}{160}\min\{D, \rho\Delta a\} = \frac{\min\{D, \rho\Delta a\}}{8}.$$

□

Proposition D.30 (Comparison rate specialization)**.** *Under the standing assumptions, let a finite integer N satisfy the N condition of Assumption 12. Then with $e_N = e^{-\eta\gamma N}$,*

$$e_N \leq \frac{1}{4}, \qquad B_{\mathrm{err}}e_N \leq \frac{D}{4}, \qquad \bar{G}_N - \bar{G}_0 \geq \frac{3\rho\Delta a}{4}, \qquad \bar{F}_0 - \bar{F}_N \geq \frac{3D}{4}.$$

The terminal clock bound also implies $a_* - a(N) \leq \rho\Delta a e^{-\lambda_a H_N} \leq \rho\Delta a e_N$, retaining both lower bounds on H_N.

Proof. The numerical condition gives

$$\eta\gamma N \geq \log\big(\max\{4, 4B_{\mathrm{err}}/D\}\big), \quad 0 < e_N \leq \frac{1}{\max\{4, 4B_{\mathrm{err}}/D\}} \leq \min\left\{\frac{1}{4}, \frac{D}{4B_{\mathrm{err}}}\right\}.$$

The clock bounds imply

$$H_N \geq \eta\sum_{n=0}^{N-1}\sigma(\mathrm{logit}(c_0) - \eta n\rho\Delta c)^d, \qquad H_N \geq \eta N c_*^d.$$

$$\lambda_a H_N \geq \eta N\lambda_a c_*^d \geq \eta N\gamma, \qquad e^{-\lambda_a H_N} \leq e_N,$$

The contraction and residual estimates therefore yield

$$\begin{aligned}\bar{G}_N - \bar{G}_0 = a(N) - a_0 = \rho\Delta a - (a_* - a(N)) \\ \geq \rho\Delta a - \rho\Delta a e_N \geq \rho\Delta a - \frac{\rho\Delta a}{4} = \frac{3\rho\Delta a}{4}. \\ \bar{F}_0 - \bar{F}_N = A_0 - \bar{F}_N = D - (\bar{F}_N - A_*) \\ \geq D - |\bar{F}_N - A_*| \geq D - B_{\mathrm{err}} e_N \geq D - \frac{D}{4} = \frac{3D}{4}.\end{aligned}$$

□

Proposition D.31 (Finite-network separation)**.** *Under the setting's Assumptions 7, 8, 9, 10, and 11, for every finite N, L satisfying Assumption 12 after Lemma D.26, Proposition D.30 and Proposition D.28 imply*

$$G_L(\theta_N) - G_L(\theta_0) \geq \frac{\rho\Delta a}{2} > 0, \qquad F_L(\theta_0) - F_L(\theta_N) \geq \frac{D}{2} > 0.$$

The final tied-coordinate and four endpoint discrepancy bounds remain bounded by the prescribed $E_N(L)$.

Proof. Write $e_N = e^{-\eta\gamma N}$. Retaining both endpoint errors gives

$$\begin{aligned}G_L(\theta_N) - G_L(\theta_0) &\geq \bar{G}_N - \bar{G}_0 - \delta_{G,N} - \delta_{G,0} \\ &\geq \rho\Delta a(1 - e_N) - 2E_N(L), \\ F_L(\theta_0) - F_L(\theta_N) &\geq \bar{F}_0 - \bar{F}_N - \delta_{F,0} - \delta_{F,N} \\ &\geq D - B_{\mathrm{err}} e_N - 2E_N(L).\end{aligned}$$

The finite-budget condition implies

$$\delta_{G,N} + \delta_{G,0} \leq 2E_N(L) \leq \frac{\min\{D, \rho\Delta a\}}{4} \leq \frac{\rho\Delta a}{4},$$

$$\delta_{F,0} + \delta_{F,N} \leq 2E_N(L) \leq \frac{\min\{D, \rho\Delta a\}}{4} \leq \frac{D}{4}.$$

Using $e_N \leq 1/4$ and $B_{\mathrm{err}} e_N \leq D/4$,

$$G_L(\theta_N) - G_L(\theta_0) \geq \rho\Delta a - \frac{\rho\Delta a}{4} - \frac{\rho\Delta a}{4} = \frac{\rho\Delta a}{2},$$

$$F_L(\theta_0) - F_L(\theta_N) \geq D - \frac{D}{4} - \frac{D}{4} = \frac{D}{2}.$$

Both margins are positive. The parameter and endpoint discrepancy bounds are those of Proposition D.28. □

Proposition D.32 (Strict finite-teacher reference advantage)**.** *Under the standing assumptions, let finite N, L satisfy the L condition of Assumption 12. Then the same frozen finite teacher's actual free accuracies satisfy*

$$T_L(Y) - T_L(\perp) \geq \Delta a - 2E_N(L) \geq (1 - \rho/4)\Delta a > 0.$$

Proof. First,

$$A_0 = u_0 + (a_0 - u_0)c_0^d \leq u_0 + (a_0 - u_0) = a_0,$$

The two teacher coupling bounds give

$$\begin{aligned}T_L(Y) - T_L(\perp) &\geq a_{\mathrm{ref}} - A_0 - |T_L(Y) - a_{\mathrm{ref}}| - |T_L(\perp) - A_0| \\ &\geq a_{\mathrm{ref}} - a_0 - 2K^2 M e^{-L} = \Delta a - 2K^2 M e^{-L}.\end{aligned}$$

Since

$$\frac{K^2 M e^{-L}}{E_N(L)} = \frac{1}{20K^2(N+1)e^{R_N}(1 + 4\eta K^3)^N} \leq \frac{1}{80} \leq 1,$$

$$2K^2 M e^{-L} \leq 2E_N(L) \leq \frac{\min\{D, \rho\Delta a\}}{4} \leq \frac{\rho\Delta a}{4},$$

the lower bound is at least $(1 - \rho/4)\Delta a > 0$. □

D.2.7 EXACT ZERO-PRIVILEGE CHECKPOINT RECOVERY

Lemma D.27 (Finite differentiable prefix law)**.** *Under the standing assumptions, fix any capped prediction prefix h, either reference value, and any finite $\theta \in \mathbb{R}^K$. The actual output logits are finite affine functions of θ, hence the full softmax is differentiable in θ and strictly positive on every token of $\mathcal{V}$.*

Proof. At a fixed finite prefix, the frozen representation is finite and independent of θ; every logit is affine in θ. Thus

$$p_{\theta,L}(v \mid h, \perp) = \frac{\exp(\ell_v(\theta; h))}{\sum_{u \in \mathcal{V}} \exp(\ell_u(\theta; h))} > 0.$$

The positive finite denominator gives differentiability. Literal continuations always index a valid table, so this also applies to malformed prefixes. □

Lemma D.28 (Pointwise stopped-KL cancellation)**.** *Under the standing assumptions, for every capped prediction prefix h, if $B = 0$, then*

$$\nabla_\theta \operatorname{KL}\big(q_{T,L}(\cdot \mid h, 0) \| p_{\theta,L}(\cdot \mid h, \perp)\big)\big|_{\theta=\theta_0} = 0 \in \mathbb{R}^K.$$

This includes every incorrect-state or out-of-order prefix, irrespective of its probability under a rollout.

Proof. At $B = 0$, the common checkpoint gives

$$q_{T,L}(v \mid h, 0) = p_{\theta_0,L}(v \mid h, \perp) \quad (v \in \mathcal{V}).$$

For fixed q_v and prefix, the log-softmax derivative is

$$\partial_{\theta_k} \log p_{\theta,L}(v \mid h, \perp) = \partial_{\theta_k} \ell_v(\theta; h) - \sum_{u \in \mathcal{V}} p_{\theta,L}(u \mid h, \perp) \partial_{\theta_k} \ell_u(\theta; h).$$

Therefore

$$\begin{aligned} \partial_{\theta_k} \sum_{v \in \mathcal{V}} q_v \log \frac{q_v}{p_{\theta,L}(v \mid h, \perp)} &= -\sum_{v \in \mathcal{V}} q_v \partial_{\theta_k} \ell_v(\theta; h) + \sum_{u \in \mathcal{V}} p_{\theta,L}(u \mid h, \perp) \partial_{\theta_k} \ell_u(\theta; h) \\ &= \sum_{v \in \mathcal{V}} \big(p_{\theta,L}(v \mid h, \perp) - q_v\big) \partial_{\theta_k} \ell_v(\theta; h). \end{aligned}$$

At θ_0, every difference vanishes, including all suppressed-row probabilities. □

Proposition D.33 (Exact finite-checkpoint invariance)**.** *Under the standing assumptions, the actual population iteration is well-defined and satisfies $\theta_n = \theta_0$ for every integer $n \geq 0$.*

Proof. At $\rho = 0$, $B = 0$ almost surely. If $\theta_n = \theta_0$, pointwise cancellation at every visited prefix gives

$$\theta_{n+1} = \theta_0 - \eta\, \mathbb{E}\left[\sum_{h \in \mathcal{H}(\tau)} 0\right] = \theta_0.$$

Starting from the common initialization, induction proves the identity for every finite n. □

Proposition D.34 (Exact free and position-forced recovery)**.** *Under the standing assumptions, for each integer $n \geq 0$,*

$$F_L(\theta_n) = F_L(\theta_0), \qquad G_L(\theta_n) = G_L(\theta_0).$$

Proof. Parameter equality gives identical token kernels at every prefix, hence identical stopped free-string probabilities. The forced kernel at positions $m \leq d$ is

$$\frac{p_{\theta,L}(C_{j,i} \mid h, \perp)}{\sum_{a=1}^{d} \sum_{b=1}^{M} p_{\theta,L}(C_{a,b} \mid h, \perp)}.$$

Its denominator is positive and identical at $\theta_n = \theta_0$; the ordinary final-position kernels also agree. Products of these kernels give identical forced-string probabilities. Summing the same success indicator proves both accuracy equalities. □

D.3 THEOREM 3

Throughout this section, we work under the setting in Appendix B.6.

D.3.1 EXACT NEURAL REALIZATION AND A SHARED CHECKPOINT

Lemma D.29 (Parallel literal retrieval)**.** *Under the standing assumptions, there are explicit frozen embeddings, query/key/value maps for two parallel heads, and W_O such that, for every table collection, literal unanswered query $(t, s) \in \{0, \dots, n\} \times [M]$, and reference $R \in \{\perp, Y\}$, the two last-position head outputs are exactly $e_{f_{t+1}(s)}$ and (B, Be_Y). The head-output and reserved-indicator fields start at zero; W_O writes only the head-output fields and leaves the stage and constant fields unchanged. All selections use only causally visible input embeddings, including for incorrect or nonconsecutive literal continuation tokens.*

Proof. Use disjoint residual blocks:

Block	Dimension	Coordinates
const	1	constant
tk	$K + M$	table lookup-stage and input-state key
tv	M	table value
gq	$K + M$	generated lookup-stage and current-state query
stage	K	literal prediction stage
mark	1	reference-slot marker
rv	$1 + M$	reference indicator and answer value
so	M	step-head output
ro	$1 + M$	reference-head output
ind	$KM + KM^2$	separate $(0, t, z)$ and $(1, t, z, y)$ reserved slots

Let P_D, I_D be their coordinate projections and injections, so $P_D I_D = \mathrm{Id}$ and $P_D I_{D'} = 0$ for $D \neq D'$. Unspecified entries are zero. Embed table rows as

$$I_{\mathrm{const}} 1 + I_{\mathrm{tk}}(e_\ell \oplus e_i) + I_{\mathrm{tv}} e_{f_\ell(i)}.$$

initial queries as

$$I_{\mathrm{const}} 1 + I_{\mathrm{gq}}(e_1 \oplus e_s) + I_{\mathrm{stage}} e_0.$$

literal continuations $C_{j,i}$ as

$$I_{\mathrm{const}} 1 + I_{\mathrm{gq}}(e_{j+1} \oplus e_i) + I_{\mathrm{stage}} e_j.$$

and the reference slot as

$$I_{\mathrm{const}} 1 + I_{\mathrm{mark}} 1 + I_{\mathrm{rv}}(B, Be_Y),$$

with $(B, Be_Y) = 0$ at empty reference. Other tokens have only constant coordinate one. The causal mask exposes all tables and the reference slot. Choose

$$W_Q^{\mathrm{step}} = P_{\mathrm{gq}}, \qquad W_K^{\mathrm{step}} = P_{\mathrm{tk}}, \qquad W_V^{\mathrm{step}} = P_{\mathrm{tv}}.$$

The table-row score is

$$(e_{t+1} \oplus e_s)^\top (e_\ell \oplus e_i) = \mathbf{1}\{\ell = t + 1\} + \mathbf{1}\{i = s\}.$$

so $(t + 1, s)$ uniquely scores two; competitors score at most one. Hard attention returns $e_{f_{t+1}(s)}$. Choose

$$W_Q^{\mathrm{ref}} = P_{\mathrm{const}}, \qquad W_K^{\mathrm{ref}} = P_{\mathrm{mark}}, \qquad W_V^{\mathrm{ref}} = P_{\mathrm{rv}}.$$

The reference slot uniquely scores one and returns (B, Be_Y). Both heads read the original embeddings. With

$$W_O[v; w] = I_{\mathrm{so}} v + I_{\mathrm{ro}} w.$$

the first residual satisfies

$$P_{\mathrm{const}} h = 1, \quad P_{\mathrm{stage}} h = e_t, \quad P_{\mathrm{so}} h = e_z, \quad P_{\mathrm{ro}} h = (B, Be_Y), \quad P_{\mathrm{ind}} h = 0.$$

Thus lookup is exact after every literal continuation, including an incorrect one. □

Lemma D.30 (Residual indicator isolation)**.** *Under the standing assumptions, there are explicit frozen W_1, W_2 of MLP width $KM(M+1)$ and residual dimension $KM(M+1)+3K+6M+4$ such that the reserved coordinates of $\widetilde{h} = h + W_2\,\mathrm{ReLU}(W_1 h)$ are exactly the setting's disjoint ϕ^0, ϕ^1 on every literal unanswered query. At $B=0$ precisely the coordinate $(0,t,z)$ is 1; at $B=1$ precisely $(1,t,z,Y)$ is 1. If the readout vanishes on all nonreserved coordinates as prescribed, its logits are exactly this one reserved column, regardless of any raw residual contents.*

Proof. For the actual stage/value (τ,ζ), write

$$a_t = \mathbf{1}\{t=\tau\}, \qquad b_z = \mathbf{1}\{z=\zeta\}, \qquad r_y = B\mathbf{1}\{y=Y\}, \qquad P_{\mathrm{const}}h = 1.$$

Define the hidden rows by

$$(W_1h')_{0,t,z} = (P_{\mathrm{stage}}h')_t + (P_{\mathrm{so}}h')_z - (P_{\mathrm{ro}}h')_B - P_{\mathrm{const}}h',$$
$$(W_1h')_{1,t,z,y} = (P_{\mathrm{stage}}h')_t + (P_{\mathrm{so}}h')_z + (P_{\mathrm{ro}}h')_B + (P_{\mathrm{ro}}h')_y - 3P_{\mathrm{const}}h'.$$

Their ReLU outputs satisfy

$$[a_t + b_z - B - 1]_+ = \mathbf{1}\{B=0, t=\tau, z=\zeta\},$$
$$[a_t + b_z + B + r_y - 3]_+ = \mathbf{1}\{B=1, t=\tau, z=\zeta, y=Y\}.$$

Indeed the matching binary sum gives one, and every mismatch gives a nonpositive input. Let W_2 copy these outputs to the initially zero indicator block. The width is $KM + KM^2$, and the residual dimension is

$$1 + (K+M) + M + (K+M) + K + 1 + (1+M) + M + (1+M) + KM(M+1)$$
$$= KM(M+1) + 3K + 6M + 4.$$

Assign arbitrary logit columns to reserved coordinates, with all raw coefficients zero. Then

$$U\widetilde{h} = \sum_{t,z} \phi^0_{t,z} L^0_{t,z} + \sum_{t,z,y} \phi^1_{t,z,y} L^1_{t,z,y}.$$

selects exactly $L^0_{\tau,\zeta}$ or $L^1_{\tau,\zeta,Y}$. The no-reference columns are independent and depend only on $F=(t,z)$. □

Lemma D.31 (Independent probability columns in the closure)**.** *Under the standing assumptions, every assignment $v^0_{t,z} \in \Delta(\mathcal{V})$ and $v^1_{t,z,y} \in \Delta(\mathcal{V})$ is simultaneously represented in the stipulated closure. Holding the reference assignments v^1 fixed, the no-reference policies in Π_0 have precisely the freedom $\prod_{t=0}^{n}\prod_{z=1}^{M}\Delta(\mathcal{V})$, with outputs evaluated at the actual feature $F=(t,z)$ and all raw readout coefficients zero.*

Proof. For any target column v, put

$$v_\varepsilon(o) = (1-\varepsilon)v(o) + \frac{\varepsilon}{KM}, \qquad L_\varepsilon(o) = \log v_\varepsilon(o).$$

Then $v_\varepsilon(o) \geq \varepsilon/(KM) > 0$, $\sum_o v_\varepsilon(o) = 1$, and

$$\mathrm{softmax}(L_\varepsilon)(o) = \frac{\exp(\log v_\varepsilon(o))}{\sum_{o'}\exp(\log v_\varepsilon(o'))} = \frac{v_\varepsilon(o)}{1} = v_\varepsilon(o).$$

$$|v_\varepsilon(o) - v(o)| = \varepsilon\left|\frac{1}{KM} - v(o)\right| \leq \varepsilon,$$

Apply this simultaneously to the finitely many reserved columns, keeping reference targets fixed. Conversely every limiting column is nonnegative and sums to one. Hence the closure is exactly the stated product of simplices; raw readout coefficients remain zero. □

Proposition D.35 (A common controlled neural checkpoint)**.** *Under the standing assumptions, there is one closure checkpoint θ_0 whose outputs are exactly the prescribed q_0, q_1 on every literal unanswered input, for every admissible pair of arbitrary column-stochastic matrices Q_0, Q_1. Four copies of it have the identical initial student policy $p_0 = q_0$. Freezing its reference columns and all its non-readout maps leaves exactly the no-reference freedom of Π_0, with every raw readout coefficient still zero.*

Proof. Assign, for $t < n$,

$$v^0_{t,z} = c_0\delta_{C_{t+1,z}} + (1-c_0)\sum_{i=1}^{M} u\delta_{A_i},$$

$$v^1_{t,z,y} = c_1\delta_{C_{t+1,z}} + (1-c_1)\sum_{i=1}^{M} Q_1(i \mid y)\delta_{A_i}.$$

and terminally,

$$v^0_{n,z} = \sum_{i=1}^{M} Q_0(i \mid z)\delta_{A_i}, \qquad v^1_{n,z,y} = \sum_{i=1}^{M} Q_1(i \mid y)\delta_{A_i}.$$

Nonnegativity and column stochasticity give

$$c_0 + (1-c_0)Mu = 1, \qquad c_1 + (1-c_1)\sum_{i=1}^{M} Q_1(i \mid y) = 1.$$

and terminal sums one. The preceding lemmas realize these columns with the same frozen maps. Literal retrieval selects $z = f_{t+1}(s)$, giving exactly q_0, q_1, including malformed prefixes. Use this single checkpoint for the teacher and all student initializations, and freeze its reference columns thereafter. □

D.3.2 Population projection and self-generated support

Lemma D.32 (Finite relative entropy and its equality case)**.** *For arbitrary probability vectors $a, b \in \Delta(\mathcal{V})$, under the setting's extended-relative-entropy convention, $D_{\mathrm{KL}}(a\|b) \geq 0$, and equality holds if and only if $a = b$. With $H(a) = -\sum_{v:a(v)>0} a(v)\log a(v)$, one has $0 \leq H(a) \leq \log(KM)$. No setting randomness or fitting assumption is required.*

Proof. Let $S = \{v : a(v) > 0\}$. If $a(v) > 0 = b(v)$, KL is infinite. Otherwise $-\log x \geq 1 - x$ gives

$$D_{\mathrm{KL}}(a\|b) \geq \sum_{v\in S}(a(v) - b(v)) = 1 - \sum_{v\in S} b(v) = \sum_{v\notin S} b(v) \geq 0.$$

Equality requires $a(v) = b(v)$ on S and zero b-mass outside S, hence $a = b$. Entropy is nonnegative, and

$$0 \leq D_{\mathrm{KL}}(a\|\,\mathrm{Unif}(\mathcal{V})) = \log(KM) - H(a),$$

gives its upper bound. □

Lemma D.33 (Token-occupancy KL decomposition)**.** *Under the independent-column domain of Lemma D.31 and the setting's finite collection and loss definitions, fix any collection/label law μ with at most K collected unanswered prefixes and simplex label q at each prefix. Let*

$$d_f = \mathbb{E}_\mu \sum_{\xi \text{ collected}} \mathbf{1}\{F(\xi) = f\}, \qquad m_f(v) = \frac{\mathbb{E}_\mu \sum_{\xi \text{ collected}} \mathbf{1}\{F(\xi) = f\}q(v)}{d_f} \quad (d_f > 0).$$

Choose any simplex column at $d_f = 0$, so that $m \in \Pi_0$. For every $r \in \Pi_0$, with $r_f = r(\cdot \mid f)$,

$$J_P(r;\mu) = C(\mu) + \sum_{f:d_f>0} d_f D_{\mathrm{KL}}(m_f\|r_f), \qquad C(\mu) = J_P(m;\mu) \in [0, K\log(KM)].$$

This is an extended-real equality and contains no product $0 \cdot \infty$. Its minimizers are exactly the policies satisfying $r_f = m_f$ for every occupied f; unoccupied columns are arbitrary.

Proof. Index the finitely many collected occurrences by a, with outcome weight $\lambda_a > 0$, feature f_a, and label q_a. Thus

$$\sum_a \lambda_a = \mathbb{E}_\mu[\#\{\text{collected prefixes}\}] \leq K, \quad d_f = \sum_{a:f_a=f} \lambda_a, \quad d_f m_f(v) = \sum_{a:f_a=f} \lambda_a q_a(v).$$

In particular m_f is a probability vector; $m_f(v) = 0$ implies that all positive-weight labels at f vanish there. If r_f is positive on the mean support, finite grouping gives

$$J_P(r;\mu) = \sum_a \lambda_a \sum_{v:q_a(v)>0} q_a(v)\log q_a(v) - \sum_{f:d_f>0} d_f \sum_{v:m_f(v)>0} m_f(v)\log r_f(v),$$
$$J_P(m;\mu) = \sum_a \lambda_a \sum_{v:q_a(v)>0} q_a(v)\log q_a(v) - \sum_{f:d_f>0} d_f \sum_{v:m_f(v)>0} m_f(v)\log m_f(v).$$

Subtracting yields the decomposition, with

$$C(\mu) = \sum_{f:d_f>0} d_f H(m_f) - \sum_a \lambda_a H(q_a).$$

$$0 \le C(\mu) \le \sum_{f:d_f>0} d_f \log(KM) \le K\log(KM).$$

If $m_f(v) > 0 = r_f(v)$ at an occupied feature, both sides are $+\infty$, while $C(\mu)$ remains finite. Independent columns make m feasible. Nonnegativity and the equality condition of KL show that minimizers equal m_f exactly on occupied features; null columns are free. □

Lemma D.34 (Occupied learner columns have prescribed support and hazards)**.** *Under the standing assumptions, fix $P \in \{\mathrm{ON}, \mathrm{PROCESS}\}$ and a policy $p \in \Pi_0$ satisfying the corresponding premise of Assumption 17. Compute d_f using its own frozen law $\mu_P[p]$. If $d_{(t,z)} > 0$ and $t < n$, then*

$$p(\mathcal{V} \setminus (\{C_{t+1,z}\} \cup \{A_i : i \in [M]\}) \mid (t,z)) = 0, \qquad p(C_{t+1,z} \mid (t,z)) = c_P,$$

where $c_{\mathrm{ON}} = \bar{c} = (1-\rho)c_0 + \rho c_1$ and $c_{\mathrm{PROCESS}} = c_0$. If $d_{(n,z)} > 0$, then $p(\{A_i : i \in [M]\} \mid (n,z)) = 1$. Both constants satisfy $0 < c_P \le 1$.

Proof. Freeze $\mu_P[p]$. The projection lemma gives $p_f = m_f$ at occupied features. Each label is supported on the designated continuation and answers at unfinished features, and only answers terminally; its mean has the same restriction. For ON, let N_f be the feature count. The single labeling bit is independent of the collection, so

$$\mathbb{E}_{\mu_{\mathrm{ON}}[p]} \sum_{\xi \text{ collected}} \mathbf{1}\{F(\xi) = f\} q_B(C_{t+1,z}) = \mathbb{E}[N_f c_B] = \mathbb{E}[N_f]\big((1-\rho)c_0 + \rho c_1\big) = d_f \bar{c}.$$

Dividing by $d_f > 0$ gives continuation mass $\bar{c}$. PROCESS labels unfinished features with q_0, giving mass c_0. Finally,

$$0 < (1-\rho)c_0 \le \bar{c} \le c_0 \le 1.$$

ensures positivity. □

Proposition D.36 (No first malformed continuation)**.** *Under the standing assumptions, fix any ON/PROCESS policy satisfying the same explicit premise of Assumption 17. Under its actual empty-reference evaluation law, almost surely every visited prediction feature has positive own-law training occupancy; every emitted continuation is precisely the next correct stage/state token; and an answer is emitted by token K. Thus an unanswered prefix after j emitted tokens, for $0 \le j \le n$, coincides almost surely with E_j, and its actual feature is $F_j = (j, S_{j+1})$.*

Proof. Collection and evaluation have the same student task/token marginal. Thus

$$\Pr_p(N_f \ge 1) \le \mathbb{E}_p N_f = d_f.$$

A finite union shows that null-occupancy features are never queried almost surely. At each occupied feature the preceding lemma permits only the designated continuation or an answer. If a first incorrect continuation occurred after j correct ones, the feature would be $(j, f_{j+1}(S_j)) = (j, S_{j+1})$, whose only continuation is $C_{j+1,S_{j+1}}$, a contradiction. After n continuations the terminal feature permits only an answer. Hence every surviving prefix is correct and $\tau \le n+1 = K$. □

Lemma D.35 (Taskwise survival induction)**.** *Under the standing assumptions, fix an ON/PROCESS policy satisfying its premise of Assumption 17. For every task value x with $\Pr(X = x) > 0$, and simultaneously for all $0 \le j \le n$,*

$$\Pr_p(E_j \mid X = x) = c_P^j > 0, \qquad d_{(j,S_{j+1}(x))} \ge \Pr(X = x)c_P^j > 0.$$

Here $S_{j+1}(x)$ is the setting-defined composed state evaluated on that fixed task. For $j < n$, conditional on $X = x, E_j$, the next continuation has probability c_P and is $C_{j+1,S_{j+1}(x)}$; at $j = n$ the next token is an answer with probability one. No conditional law of Y is assumed.

Proof. Fix a supported task atom x. At entry,

$$d_{(0,S_1(x))} \ge \Pr(X = x) = \Pr(X = x)c_P^0 > 0.$$

so projection supplies continuation mass $c_P > 0$. Suppose $\Pr(E_j \mid X = x) = c_P^j$. The queried feature on this event has

$$d_{(j,S_{j+1}(x))} \ge \Pr_p(X = x, E_j) = \Pr(X = x)c_P^j > 0.$$

and therefore continuation mass c_P. Fresh sampling gives

$$\Pr_p(E_{j+1} \mid X = x) = \Pr_p(E_j \mid X = x)\ \Pr_p(C_{j+1,S_{j+1}(x)} \text{ next} \mid X = x, E_j) = c_P^{j+1}.$$

This also supplies the next positive occupancy. Induction proves both claims through $j = n$, where answer mass is one. □

D.3.3 Fresh suffix information under selected prefixes

Lemma D.36 (A fresh final permutation hides the unresolved answer)**.** *Under the standing assumptions, for every $0 \le j \le n$, $\Pr(S_{j+1} = z) = 1/M$ for each $z \in [M]$. For $j < n$, define the restricted task history $H_j = (S_0, f_1, \ldots, f_{j+1})$. For every positive-probability value h of this history and every $y \in [M]$,*

$$\Pr(Y = y \mid H_j = h) = 1/M, \qquad \Pr(S_{j+1} = z, Y = y) = 1/M^2.$$

At $j = n$, $S_{j+1} = Y$ identically; thus the joint mass is $\mathbf{1}\{y = z\}/M$.

Proof. Counting permutations gives

$$\Pr(f_{j+1}(S_j) = z \mid S_0, f_1, \ldots, f_j) = \frac{(M-1)!}{M!} = \frac{1}{M}.$$

so every S_{j+1} is uniform. For $j < n$, the independent final permutation gives

$$\Pr(Y = y \mid S_0, f_1, \ldots, f_{K-1}) = \Pr(f_K(S_{K-1}) = y \mid S_0, f_1, \ldots, f_{K-1}) = 1/M.$$

Averaging over unread intermediate permutations preserves this law conditional on H_j. Hence

$$\Pr(S_{j+1} = z, Y = y) = \frac{1}{M}\Pr(S_{j+1} = z) = 1/M^2.$$

At $j = n$, use $S_{j+1} = Y$ instead. □

Lemma D.37 (Teacher continuation likelihood)**.** *Under the standing assumptions, fix $b \in \{0, 1\}$. Write $\Pr_b$ for the setting's teacher sampling experiment with the collection bit fixed to b, retaining the fresh task law. For each task x and each $0 \le j \le n$,*

$$\Pr_b(E_j \mid X = x) = c_b^j.$$

Every unanswered prefix lies on the correct consecutive path almost surely; it is at depth j exactly on E_j, with feature (j, S_{j+1}). Terminal sampling is an answer. These conclusions include $c_b = 0$, interpreting its depth-zero power as 1.

Proof. Literal support allows only the next true continuation or an answer, so induction gives correct prefixes and terminal answering. For a fixed task,

$$\Pr_b(E_{j+1} \mid X = x) = \mathbb{E}_b[\mathbf{1}_{E_j} c_b \mid X = x] = c_b \Pr_b(E_j \mid X = x).$$

Starting from $E_0 = \Omega$, this yields $\Pr_b(E_j \mid X = x) = c_b^j$. No division by survival is used, so $c_b = 0$ is included. □

Proposition D.37 (Selected feature occupancy and suffix law)**.** *Under the standing assumptions, fix an ON/PROCESS learner p, or a fixed-bit teacher kernel $\Pr_b$. Then, for every $j \le n, z \in [M]$,*

$$\Pr_p(E_j, F_j = (j,z)) = d_{j,z}^{\text{learner}} = c_P^j/M > 0,$$
$$\Pr_b(E_j, F_j = (j,z)) = d_{j,z}^{(b)} = c_b^j/M,$$
$$d_{j,z}^{\text{teacher}} = D_j/M \ge (1-\rho)c_0^j/M > 0.$$

Here each d is the actual expected token count for feature (j,z) in its specified collection law. For $j < n$, the joint masses including $Y = y$ equal respectively $c_P^j/M^2, c_b^j/M^2, D_j/M^2$. For every positive selected restricted history $E_j, H_j = h$, Y remains uniform under learner collection, fixed-bit teacher collection, or mixed teacher collection. This entails uniformity conditional on the actual feature and surviving generated history. At $j = n$, $Y = z$ on $E_j, F_j = (j,z)$, and the corresponding joint masses are the displayed feature masses times $\mathbf{1}\{y = z\}$.

Proof. For any task event G, taskwise survival gives

$$\Pr_p(E_j, G) = \sum_{x \in G} \Pr(X = x) \Pr_p(E_j \mid X = x) = c_P^j \Pr(G).$$

$$\Pr_b(E_j, G) = c_b^j \Pr(G).$$

Independence of the collection bit gives

$$\Pr(E_j, G, B_{\text{col}} = 0) = (1-\rho)c_0^j \Pr(G), \qquad \Pr(E_j, G, B_{\text{col}} = 1) = \rho c_1^j \Pr(G),$$

Their sum is $D_j \Pr(G)$. The single-visit identity is

$$N_{j,z} = \mathbf{1}\{E_j, S_{j+1} = z\}.$$

Taking $G = \{S_{j+1} = z\}$ or $G = \{S_{j+1} = z, Y = y\}$ and applying the suffix lemma proves the stated occupancy and joint laws. Their denominators are positive because

$$\bar{c} = (1-\rho)c_0 + \rho c_1 \ge (1-\rho)c_0 > 0, \qquad D_j \ge (1-\rho)c_0^j > 0.$$

For restricted histories and $j < n$, the same calculation gives

$$\begin{aligned}
&\Pr_p(E_j, H_j = h, Y = y) = c_P^j \Pr(H_j = h)/M, \quad \Pr_p(E_j, H_j = h) = c_P^j \Pr(H_j = h),\\
&\Pr_b(E_j, H_j = h, Y = y) = c_b^j \Pr(H_j = h)/M, \quad \Pr_b(E_j, H_j = h) = c_b^j \Pr(H_j = h),\\
&\Pr(E_j, H_j = h, Y = y) = D_j \Pr(H_j = h)/M, \quad \Pr(E_j, H_j = h) = D_j \Pr(H_j = h).
\end{aligned}$$

Divide only where the denominator is positive. The generated unanswered history and current feature are functions of H_j, so summing over compatible h preserves the uniform answer law. Terminally, $Y = z$. □

Lemma D.38 (Original collection-bit posterior)**.** *Under the three primitive assumptions of Proposition D.37 and that proposition, in the mixed teacher collection experiment, for all $0 \le j \le n, z \in [M]$,*

$$\Pr(B_{\text{col}} = 1 \mid E_j, F_j = (j,z)) = \frac{\rho c_1^j}{D_j} = w_j.$$

For $j < n$, the conditional joint law of B_{col} and Y at that selected feature is

$$\Pr(B_{\text{col}} = b, Y = y \mid E_j, F_j = (j,z)) = \begin{cases} (1-w_j)/M, & b = 0, \\ w_j/M, & b = 1. \end{cases}$$

The same posterior and product law hold conditional on a positive selected restricted history $E_j, H_j = h$. At terminal, replace $1/M$ in the joint answer law by $\mathbf{1}\{y = z\}$. For $j < n$,

$$w_{j+1} = \frac{w_j c_1}{(1-w_j)c_0 + w_j c_1}, \qquad (1-w_j)c_0 + w_j c_1 = D_{j+1}/D_j > 0.$$

In particular, $w_0 = \rho$; if $\rho = 0$, all $w_j = 0$; if $c_1 = 0$, then $w_j = 0$ for every $j \geq 1$.

Proof. Bayes' rule and the selected joint laws give

$$w_j = \frac{\rho c_1^j / M}{D_j / M}, \qquad \Pr(B_{\mathrm{col}} = 1, Y = y \mid E_j, F_j = (j,z)) = \frac{w_j}{M} \quad (j < n).$$

The bit-zero and restricted-history laws follow by the same division; terminally $Y = z$. Furthermore,

$$(1-w_j)c_0 + w_j c_1 = \frac{(1-\rho)c_0^{j+1} + \rho c_1^{j+1}}{D_j} = \frac{D_{j+1}}{D_j} > 0.$$

$$\frac{w_j c_1}{(1-w_j)c_0 + w_j c_1} = \frac{\rho c_1^{j+1}/D_j}{D_{j+1}/D_j} = \frac{\rho c_1^{j+1}}{D_{j+1}} = w_{j+1}.$$

Thus $w_0 = \rho$; at $\rho = 0$, all $w_j = 0$, and at $c_1 = 0$, all positive-depth $w_j = 0$. Each $D_j > 0$. □

Lemma D.39 (Independent labeling bits survive selection)**.** *Under the standing assumptions, the ON labeling bit B and RELABEL labeling bit B_{lab} each have conditional probability ρ of being 1 given any positive selected feature $E_j, F_j = (j,z)$, for every $j \leq n$. They remain independent of the conditional answer: at $j < n$, their joint law with $Y = y$ is ρ/M for bit 1 and $(1-\rho)/M$ for bit 0. This holds as well for positive selected restricted histories, and for RELABEL even if the original collection bit is included whenever the conditioning event is positive. At terminal the answer law is $Y = z$. The PROCESS labeling bit has these independence properties, but its effective label weight is 0 at unfinished stages and ρ at terminal. Thus the actual label weights are exactly the setting's candidate $h_{P,j}$.*

Proof. For ON, the labeling bit is independent of the entire collection. For every task/token event G,

$$\Pr(B = 1, G) = \rho \Pr(G), \qquad \Pr(B = 0, G) = (1-\rho)\Pr(G).$$

For RELABEL, the new bit is also independent of the original collection bit:

$$\Pr(B_{\mathrm{lab}} = 1, G) = \rho \Pr(G), \qquad \Pr(B_{\mathrm{lab}} = 0, G) = (1-\rho)\Pr(G).$$

Apply these identities to selected feature/history events and their intersections with $\{Y = y\}$. The selected uniform law gives effective weight ρ in both arms. PROCESS uses q_0 unfinished and independent-bit labels terminally, giving weights $0, \rho$. RETAIN uses the original bit and thus weight w_j. These are joint identities, including zero-weight components. □

D.3.4 Exact fitted columns and existence

Lemma D.40 (Exact label means and feasible protocol columns)**.** *Under the standing assumptions, define for each P, j, z the following vector on the setting vocabulary:*

$$\begin{aligned} v_{P,j,z} &= \big((1-h_{P,j})c_0 + h_{P,j}c_1\big)\delta_{C_{j+1,z}} \\ &\quad + \sum_{i=1}^{M} \big((1-h_{P,j})(1-c_0)u + h_{P,j}(1-c_1)\overline{Q}_1(i)\big)\delta_{A_i}, \qquad j < n, \\ v_{P,n,z} &= \sum_{i=1}^{M} \big((1-h_{P,n})Q_0(i \mid z) + h_{P,n}Q_1(i \mid z)\big)\delta_{A_i}. \end{aligned} \tag{D.3.4.1}$$

All unlisted vocabulary coordinates are zero. In a teacher-collected law this is the label mean at every (j,z). In a law collected by any ON/PROCESS policy satisfying Assumption 17, it is again the label mean at every (j,z). The values $h_{P,j}$ are exactly the quantities defined above, with their probabilistic meaning established by Lemmas D.38 and D.39. Every vector in (D.3.4.1) *is a*

simplex vector, independently of a fitting premise, and hence there is a candidate $p^*_P \in \Pi_0$ assigning $p^*_P(\cdot \mid (j,z)) = v_{P,j,z}$ to every no-reference feature, with the same frozen reference columns as θ_0.

The unfinished hazards and terminal matrices of these vectors are

P	$v_{P,j,z}(C_{j+1,z}),\ j<n$	*terminal matrix*
ON	$\bar{c}$	Q_{mix}
RETAIN	D_{j+1}/D_j	$Q_{\mathrm{ret},n}$
RELABEL	$\bar{c}$	Q_{mix}
PROCESS	c_0	Q_{mix}.

(D.3.4.2)

Proof. Positive single-visit occupancy gives

$$m^P_{j,z}(v) = \mathbb{E}[q_P(v) \mid E_j, F_j = (j,z)]. \tag{D.3.4.3}$$

At $j < n$, the effective bit/answer law is

$$\Pr(\text{effective label bit} = 0, Y = y \mid E_j, F_j = (j,z)) = \frac{1-h_{P,j}}{M},$$
$$\Pr(\text{effective label bit} = 1, Y = y \mid E_j, F_j = (j,z)) = \frac{h_{P,j}}{M}.$$

Thus continuation mass is $(1-h_{P,j})c_0 + h_{P,j}c_1$, and

$$\begin{aligned} m^P_{j,z}(A_i) &= \sum_{y=1}^{M} \frac{1-h_{P,j}}{M}(1-c_0)u + \sum_{y=1}^{M} \frac{h_{P,j}}{M}(1-c_1)Q_1(i \mid y) \\ &= (1-h_{P,j})(1-c_0)u + h_{P,j}(1-c_1)\frac{1}{M}\sum_{y=1}^{M} Q_1(i \mid y). \end{aligned}$$

Terminally $Y = z$, giving the stated mixture of $Q_0(\cdot \mid z)$ and $Q_1(\cdot \mid z)$. Feasibility follows from $h_{P,j} \in [0,1]$ and

$$\overline{Q}_1(i) \geq 0, \qquad \sum_{i=1}^{M} \overline{Q}_1(i) = \frac{1}{M}\sum_{y=1}^{M}\sum_{i=1}^{M} Q_1(i \mid y) = 1.$$

$$(1-h_{P,j})c_0 + h_{P,j}c_1 + (1-h_{P,j})(1-c_0)\underbrace{Mu}_{1} + h_{P,j}(1-c_1)\underbrace{\sum_{i}\overline{Q}_1(i)}_{1} = 1.$$

Terminal columns likewise sum to one. Independent-column realization defines p^*_P. For RETAIN,

$$(1-w_j)c_0 + w_j c_1 = \frac{(1-\rho)c_0^j c_0 + \rho c_1^j c_1}{D_j} = \frac{D_{j+1}}{D_j}.$$

which gives the table's depth-dependent continuation probabilities. □

Lemma D.41 (Literal candidate trajectories and survival)**.** *Under the standing assumptions, let p^*_P be the feasible candidates of Lemma D.40. For every task x and $0 \leq j \leq n$, their empty-reference sampling obeys*

$$\Pr_{p^*_P}(E_j \mid X = x) = \begin{cases} \bar{c}^j, & P = \mathrm{ON}, \mathrm{RELABEL}, \\ D_j, & P = \mathrm{RETAIN}, \\ c_0^j, & P = \mathrm{PROCESS}. \end{cases} \tag{D.3.4.4}$$

Every continuation is the correct consecutive literal token almost surely, and an answer is emitted by token K. Every correct-depth feature has positive evaluation occupancy, equal to the corresponding right-hand side of (D.3.4.4) *divided by M. For ON/PROCESS this is also their own-law collection occupancy. These statements do not assume that the candidates are fitted.*

Proof. The candidate support yields the correct-prefix invariant by induction from $(0, S_0)$, and forces an answer at depth n. Fresh sampling gives

$$\Pr_{p_P^*}(E_{j+1} \mid X = x) = v_{P,j,S_{j+1}(x)}(C_{j+1,S_{j+1}(x)}) \Pr_{p_P^*}(E_j \mid X = x).$$

ON/RELABEL multiply survival by $\bar{c}$, and PROCESS by c_0. RETAIN gives

$$\Pr_{p^*_{\mathrm{RETAIN}}}(E_j \mid X = x) = \prod_{\ell=0}^{j-1} \frac{D_{\ell+1}}{D_\ell} = \frac{D_j}{D_0} = D_j, \tag{D.3.4.5}$$

All ratios lie in $(0, 1]$, since $0 < D_{j+1} \le D_j$. The count at (j, z) equals $\mathbf{1}\{E_j, S_{j+1} = z\}$. Task-independent survival and uniform S_{j+1} therefore give occupancy equal to the corresponding survival divided by M, strictly positive. □

Proposition D.38 (Four attained frozen-law optima)**.** *Under the standing assumptions, the four candidates satisfy*

$$\begin{aligned} p_P^* &\in \arg\min_{r \in \Pi_0} J_P(r; \mu_P), && P \in \{\mathrm{RETAIN}, \mathrm{RELABEL}\}, \\ p_P^* &\in \arg\min_{r \in \Pi_0} J_P(r; \mu_P[p_P^*]), && P \in \{\mathrm{ON}, \mathrm{PROCESS}\}. \end{aligned} \tag{D.3.4.6}$$

The second minimization freezes the collection and label law induced by the displayed candidate; no fitted-policy assumption is used to compute that law.

Proof. For RETAIN/RELABEL, the fixed teacher collection has occupancy $D_j/M > 0$, and its means are exactly the candidate columns; the projection lemma proves optimality. For ON/PROCESS, generate the feasible candidate's own law first. Its proved taskwise survival yields, for $j < n$,

$$\begin{aligned} \Pr_{p_P^*}(E_j, S_{j+1} = z, Y = y) &= \sum_{x : S_{j+1}(x) = z, Y(x) = y} \Pr(X = x) c_P^j = c_P^j / M^2, \\ \Pr_{p_P^*}(E_j, S_{j+1} = z) &= c_P^j / M. \end{aligned} \tag{D.3.4.7}$$

Terminally the joint mass is $c_P^n \mathbf{1}\{y = z\}/M$. The independent labeling bit gives

$$\begin{aligned} m^{\mathrm{ON}}_{j,z}(C_{j+1,z}) &= (1-\rho)c_0 + \rho c_1, \\ m^{\mathrm{ON}}_{j,z}(A_i) &= (1-\rho)(1-c_0)u + \rho(1-c_1)\sum_{y=1}^{M} Q_1(i \mid y)/M \quad (j < n), \\ m^{\mathrm{ON}}_{n,z}(A_i) &= (1-\rho)Q_0(i \mid z) + \rho Q_1(i \mid z). \end{aligned} \tag{D.3.4.8}$$

and

$$\begin{aligned} m^{\mathrm{PROCESS}}_{j,z} &= c_0 \delta_{C_{j+1,z}} + (1-c_0)\sum_i u \delta_{A_i}, \quad j < n; \\ m^{\mathrm{PROCESS}}_{n,z}(A_i) &= (1-\rho)Q_0(i \mid z) + \rho Q_1(i \mid z). \end{aligned} \tag{D.3.4.9}$$

All other coordinates vanish. These means equal the candidate columns. Freezing this law, the projection decomposition gives

$$J_P(r; \mu) = C(\mu) + \sum_{j=0}^{n} \sum_{z=1}^{M} \frac{c_P^j}{M} D_{\mathrm{KL}}\big(v_{P,j,z} \| r(\cdot \mid (j,z))\big) \;\ge\; C(\mu).$$

with equality at $r = p_P^*$. Thus self-consistency holds without assuming prior fittedness. □

Proposition D.39 (Uniqueness of fitted evaluation behavior)**.** *Under the standing assumptions, fix any P and any policy $p_P \in \Pi_0$ satisfying its corresponding Assumption 17. Then*

$$p_P(\cdot \mid (j,z)) = v_{P,j,z} = p_P^*(\cdot \mid (j,z)) \quad \textit{for every } 0 \le j \le n,\ z \in [M]. \tag{D.3.4.10}$$

In particular all fitted policies for an arm agree on its evaluation support, generate their own correct consecutive continuation tokens, answer by the cap, and obey the survival laws (D.3.4.4)–(D.3.4.5)*. No choice of a zero-training-occupancy column or unsupported prefix can alter these conclusions.*

Proof. For RETAIN/RELABEL, every collection feature has occupancy $D_j/M > 0$. For a fitted ON/PROCESS policy, the support and survival results give occupancy $c_P^j/M > 0$ before using its answer columns. In either case, projection and the computed means imply

$$p_P(\cdot \mid (j,z)) = m_{j,z}^P = v_{P,j,z} = p_P^*(\cdot \mid (j,z)) \quad \text{for all } j,z.$$

These are all no-reference features. Products of identical conditional token kernels give identical stopped trajectory laws; candidate survival and support therefore transfer to every fitted policy. □

D.3.5 Accuracy under first-answer sampling

Lemma D.42 (Task law under student survival)**.** *Under the standing assumptions, fix either $p = p_0$ or any student $p = p_P$ satisfying its corresponding Assumption 17. Write $\Pr_p$ for its fresh-task, empty-reference evaluation law and $s_j = \Pr_p(E_j)$. Then*

Evaluated policy p	$\Pr_p(E_j \mid X = x) = s_j$, *all tasks x, $0 \le j \le n$*
p_0	c_0^j
$p_{\mathrm{ON}}, p_{\mathrm{RELABEL}}$	$\bar{c}^j$
p_{RETAIN}	D_j
p_{PROCESS}	c_0^j

For every unfinished $j < n$ and $z, y \in [M]$,

$$\Pr_p(E_j, F_j = (j,z), Y = y) = \frac{s_j}{M^2}, \qquad \Pr_p(E_j, F_j = (j,z)) = \frac{s_j}{M} > 0. \tag{D.3.5.1}$$

At terminal $F_n = (n, Y)$, and

$$\Pr_p(E_n, Y = y) = \frac{s_n}{M} > 0. \tag{D.3.5.2}$$

In particular Y is uniform conditional on an unfinished surviving feature and remains uniform across tasks that reach terminal.

Proof. Use fitted-policy equality and candidate survival; for p_0, use the bit-zero teacher law. For every task event G,

$$\Pr_p(E_j, G) = \sum_{x \in G} \Pr(X = x) \Pr_p(E_j \mid X = x) = s_j \Pr(G). \tag{D.3.5.3}$$

Taking G to specify S_{j+1}, Y and applying the suffix lemma gives the selected evaluation laws. Positivity follows from $c_0, \bar{c}, D_j > 0$. □

Lemma D.43 (Success of an unfinished student answer)**.** *Under the standing assumptions, fix one of its policies p, $j < n$, and $z \in [M]$. Let T_j be the event that the first answer is emitted at output $j+1$, and let* Succ *denote the setting's success event. On the selected feature, define the actual unnormalized answer masses*

$$b_{j,z}(i) = p(A_i \mid (j,z)), \qquad r_{j,z} = \sum_{i=1}^{M} b_{j,z}(i).$$

Then

$$\begin{aligned} \Pr_p(T_j \mid E_j, F_j = (j,z)) &= r_{j,z}, \\ \Pr_p(\mathsf{Succ} \cap T_j \mid E_j, F_j = (j,z)) &= u r_{j,z}. \end{aligned} \tag{D.3.5.4}$$

If $r_{j,z} > 0$, the conditional success given that feature and stopping is exactly u, even if the normalized answer vector is nonuniform. If $\Pr_p(T_j) > 0$, then $\Pr_p(\mathsf{Succ} \mid T_j) = u$; if premature stopping at any unfinished depth has positive probability, its conditional success is also u.

Proof. Fresh sampling depends only on the selected feature, so

$$\Pr_p(Y = y, \text{next token} = A_i \mid E_j, F_j = (j,z)) = \Pr_p(Y = y \mid E_j, F_j = (j,z))\, b_{j,z}(i) = \frac{b_{j,z}(i)}{M}. \tag{D.3.5.5}$$

Summing over all i, y gives answer mass $r_{j,z}$; summing over $i = y$ gives success mass $ur_{j,z}$. This uses no uniformity of the answer vector. If $r_{j,z} > 0$, the conditional success is u; if it is zero, both masses vanish. Summing over selected features gives

$$\Pr_p(\mathsf{Succ} \cap T_j) = u \Pr_p(T_j). \tag{D.3.5.6}$$

The disjoint union over premature depths has the same property. □

Proposition D.40 (Student first-answer accuracy)**.** *Under the standing assumptions, every fixed fitted policy subject to its original Assumption 17, and the original initial student, satisfy*

$$\begin{aligned} \mathcal{A}_0 &= u + (a_0 - u)c_0^n, \\ \mathcal{A}_{\mathrm{ON}} = \mathcal{A}_{\mathrm{RELABEL}} &= u + (a_{\mathrm{mix}} - u)\bar{c}^{\,n}, \\ \mathcal{A}_{\mathrm{RETAIN}} &= u + (1-\rho)c_0^n(a_0 - u) + \rho c_1^n(a_1 - u), \\ \mathcal{A}_{\mathrm{PROCESS}} &= u + (a_{\mathrm{mix}} - u)c_0^n. \end{aligned} \tag{D.3.5.7}$$

These are accuracies for the same sampled neural policies and the original first-answer, K-output success rule.

Proof. Correct support and terminal answering give the partition

$$T_j = E_j \setminus E_{j+1} \quad (0 \le j < n), \qquad T_n = E_n \quad \text{up to null events.} \tag{D.3.5.8}$$

Thus $\Pr(T_j) = s_j - s_{j+1}$ for $j < n$, $\Pr(T_n) = s_n$, and

$$\sum_{j=0}^{n-1}(s_j - s_{j+1}) + s_n = 1. \tag{D.3.5.9}$$

Premature success contributes $u(1 - s_n)$. For the actual terminal matrix Q,

$$\Pr_p(\mathsf{Succ} \cap T_n) = \sum_{y=1}^{M} \Pr_p(E_n, Y = y)Q(y \mid y) = \frac{s_n}{M}\sum_{y=1}^{M} Q(y \mid y). \tag{D.3.5.10}$$

Therefore

$$\mathcal{A}(p) = u(1 - s_n) + \frac{s_n}{M}\sum_y Q(y \mid y). \tag{D.3.5.11}$$

Substitute $(s_n, Q) = (c_0^n, Q_0)$ for the checkpoint, $(\bar{c}^{\,n}, Q_{\mathrm{mix}})$ for ON/RELABEL, and $(c_0^n, Q_{\mathrm{mix}})$ for PROCESS, using

$$\frac{1}{M}\sum_y [(1-\rho)Q_0(y \mid y) + \rho Q_1(y \mid y)] = (1-\rho)a_0 + \rho a_1 = a_{\mathrm{mix}}.$$

For RETAIN,

$$D_n(1 - w_n) = (1-\rho)c_0^n, \qquad D_n w_n = \rho c_1^n.$$

hence

$$\begin{aligned} \mathcal{A}_{\mathrm{RETAIN}} &= u - uD_n + (1-\rho)c_0^n a_0 + \rho c_1^n a_1 \\ &= u + (1-\rho)c_0^n(a_0 - u) + \rho c_1^n(a_1 - u), \end{aligned}$$

□

Proposition D.41 (Frozen teacher accuracy with disclosure)**.** *Under the standing assumptions, the frozen teacher with its reference fixed to Y has accuracy a_1. Under the original once-per-task disclosure mixture,*

$$\mathcal{A}_T(\rho) = (1-\rho)\mathcal{A}_0 + \rho a_1. \tag{D.3.5.12}$$

Proof. Fix the disclosed teacher kernel and a task with $Y(x) = y$. Its true-prefix survival gives

$$\Pr_1(\mathsf{Succ} \cap T_j \mid X = x) = c_1^j(1 - c_1)Q_1(y \mid y). \tag{D.3.5.13}$$

The terminal success mass is $c_1^n Q_1(y \mid y)$, so

$$Q_1(y \mid y)\left[\sum_{j=0}^{n-1} c_1^j(1 - c_1) + c_1^n\right] = Q_1(y \mid y), \tag{D.3.5.14}$$

by telescoping, including $c_1 = 0$. Averaging uniform Y gives a_1. The bit-zero kernel is p_0; mixing the two full rollout laws gives $\mathcal{A}_T = (1 - \rho)\mathcal{A}_0 + \rho a_1$. □

D.3.6 Protocol comparisons and exact recovery

Proposition D.42 (Process and teacher gains)**.** *Under the standing assumptions, let each compared student satisfy its own Assumption 17. Then*

$$\begin{aligned} \mathcal{A}_{\mathrm{PROCESS}} - \mathcal{A}_0 &= \rho(a_1 - a_0)c_0^n \geq 0, \\ \mathcal{A}_{\mathrm{PROCESS}} - \mathcal{A}_{\mathrm{ON}} &= (a_{\mathrm{mix}} - u)(c_0^n - \bar{c}^n) \geq 0, \\ \mathcal{A}_T(\rho) - \mathcal{A}_T(0) &= \rho(a_1 - \mathcal{A}_0) \geq 0. \end{aligned} \tag{D.3.6.1}$$

Proof. Subtract the exact accuracy formulas, using

$$a_{\mathrm{mix}} - a_0 = \rho(a_1 - a_0), \qquad a_{\mathrm{mix}} - u = (1 - \rho)(a_0 - u) + \rho(a_1 - u) > 0.$$

The PROCESS–ON difference has nonnegative factors because

$$c_0 - \bar{c} = \rho(c_0 - c_1) \geq 0, \qquad 0 < (1 - \rho)c_0 \leq \bar{c} \leq c_0,$$

$$c_0^n - \bar{c}^n = (c_0 - \bar{c}) \sum_{\ell=0}^{n-1} c_0^{n-1-\ell}\bar{c}^{\ell}$$

Hence the PROCESS–ON difference is strict when $\rho > 0$. The PROCESS–checkpoint difference is $\rho(a_1 - a_0)c_0^n \geq 0$. Finally,

$$a_1 - \mathcal{A}_0 = (a_1 - a_0) + (a_0 - u)(1 - c_0^n) \geq 0.$$

This gives the teacher improvement $\rho(a_1 - \mathcal{A}_0) \geq 0$. □

Proposition D.43 (Sharp source comparison)**.** *Under the standing assumptions, suppose $c_0 = 1, c_1 = 0, Q_1 = I_M$ and $0 < \rho < 1$. For any original conditionally fitted ON and RETAIN policies,*

$$\mathcal{A}_{\mathrm{ON}} - \mathcal{A}_{\mathrm{RETAIN}} = (1 - \rho)[1 - (1 - \rho)^n]\big(T_{K,M}(\rho) - a_0\big). \tag{D.3.6.2}$$

Consequently $\mathcal{A}_{\mathrm{ON}} \leq \mathcal{A}_{\mathrm{RETAIN}}$ if and only if $a_0 \geq T_{K,M}(\rho)$, and equality of the accuracies holds exactly when $a_0 = T_{K,M}(\rho)$.

Proof. Put $r = 1 - \rho \in (0, 1)$. In the specified regime $a_1 = 1, \bar{c} = r, c_1^n = 0$, so

$$\mathcal{A}_{\mathrm{ON}} = u + (ra_0 + \rho - u)r^n, \qquad \mathcal{A}_{\mathrm{RETAIN}} = u + r(a_0 - u). \tag{D.3.6.3}$$

Using $ra_0 + \rho - u = r(a_0 - u) + \rho(1 - u)$,

$$\begin{aligned} \mathcal{A}_{\mathrm{ON}} - \mathcal{A}_{\mathrm{RETAIN}} &= r^{n+1}(a_0 - u) + \rho(1 - u)r^n - r(a_0 - u) \\ &= r\{\rho(1 - u)r^{n-1} - (a_0 - u)(1 - r^n)\} \\ &= r(1 - r^n)\left\{u + \frac{\rho(1 - u)r^{n-1}}{1 - r^n} - a_0\right\}. \end{aligned} \tag{D.3.6.4}$$

The final bracket is $T_{K,M}(\rho) - a_0$. Since $r(1 - r^n) > 0$, this proves the exact sign and equality characterization. □

Lemma D.44 (Range of the source threshold)**.** *Under the setting's $K, M \geq 2$ from Assumption 13, with $0 < \rho < 1$ as allowed by Assumption 16,*

$$T_{K,M}(\rho) = u + (1-u)\frac{(1-\rho)^{n-1}}{\sum_{\ell=0}^{n-1}(1-\rho)^{\ell}}. \tag{D.3.6.5}$$

Thus $T_{2,M}(\rho) = 1$. If $K > 2$, then $u < T_{K,M}(\rho) < 1$.

Proof. For $r = 1 - \rho \in (0, 1)$,

$$(1-r)\sum_{\ell=0}^{n-1} r^{\ell} = \sum_{\ell=0}^{n-1} r^{\ell} - \sum_{\ell=1}^{n} r^{\ell} = 1 - r^{n}.$$

Cancel $1 - r = \rho > 0$ to obtain

$$T_{K,M}(\rho) = u + (1-u)\frac{r^{n-1}}{\sum_{\ell=0}^{n-1} r^{\ell}}.$$

The ratio is one for $n = 1$ and strictly between zero and one for $n > 1$. Hence $T_{2,M} = 1$ and $u < T_{K,M} < 1$ for $K > 2$. □

Proposition D.44 (Attainable source orderings)**.** *Under the source specialization $c_0 = 1$, $c_1 = 0$, $Q_1 = I_M$, and $0 < \rho < 1$, if $K = 2$ then $\mathcal{A}_{\mathrm{ON}} > \mathcal{A}_{\mathrm{RETAIN}}$ whenever $a_0 < 1$, with equality at $a_0 = 1$. For each fixed $K > 2$ and $M \geq 2$, admissible Q_0 realize either strict ordering, and an admissible Q_0 gives equality.*

Proof. For $K = 2$, the threshold identity gives

$$\mathcal{A}_{\mathrm{ON}} - \mathcal{A}_{\mathrm{RETAIN}} = (1-\rho)\rho(1-a_0) > 0$$

when $a_0 < 1$, with equality at $a_0 = 1$. For witnesses, let $H(i \mid y) = \mathbf{1}\{i = 1 + (y \bmod M)\}$, a derangement matrix, and set

$$Q_0^{(a)}(i \mid y) = a\mathbf{1}\{i = y\} + (1-a)H(i \mid y). \tag{D.3.6.6}$$

For $u < a \leq 1$,

$$\sum_{i=1}^{M} Q_0^{(a)}(i \mid y) = a + (1-a) = 1, \qquad Q_0^{(a)}(y \mid y) = a, \qquad \frac{1}{M}\sum_{y} Q_0^{(a)}(y \mid y) = a.$$

Thus $a_0 = a$, and $Q_1 = I_M, c_0 = 1, c_1 = 0$ meet the assumptions. For $K > 2$, choose

$$a_- = \frac{u + T_{K,M}(\rho)}{2}, \qquad a_+ = \frac{T_{K,M}(\rho) + 1}{2}.$$

Then $u < a_- < T_{K,M} < a_+ < 1$, producing the two strict orderings; $a = T_{K,M}$ gives equality. Existence of the corresponding fitted policies follows from Proposition D.38. □

Lemma D.45 (Identical checkpoint recovery without disclosure)**.** *Under the standing assumptions, suppose $\rho = 0$. For every arm P and every policy satisfying its original Assumption 17, the categorical column at every reachable empty-reference feature equals the initial checkpoint column q_0. For every task and each depth through the cap, the probabilities of all prefixes and stopped histories agree with those of $p_0 = q_0$. Hence all four accuracies are exactly $\mathcal{A}_0$, and $\mathcal{A}_T(0) = \mathcal{A}_0$.*

Proof. At $\rho = 0$,

$$D_j = c_0^j > 0, \qquad w_j = \frac{0 \cdot c_1^j}{c_0^j} = 0, \qquad \bar{c} = c_0, \qquad a_{\mathrm{mix}} = a_0.$$

Every effective label weight vanishes, hence

$$\begin{aligned} p_P(\cdot \mid (j,z)) &= c_0\delta_{C_{j+1,z}} + (1-c_0)\sum_i u\delta_{A_i}, \quad j < n, \\ p_P(\cdot \mid (n,z)) &= \sum_i Q_0(i \mid z)\delta_{A_i}. \end{aligned} \tag{D.3.6.7}$$

These are exactly the checkpoint columns; reference columns are already frozen. Products of identical token kernels give identical stopped laws. Consequently

$$\mathcal{A}_P = u + (a_0 - u)c_0^n = \mathcal{A}_0 \quad \text{for every } P, \qquad \mathcal{A}_T(0) = \mathcal{A}_0.$$

□