WHEN FINE-TUNING GAINS REVERSE: SEPARATING PREFIX AND CONTINUATION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) can improve accuracy at a short output budget yet reduce it at a longer one, relative to the same initial checkpoint. To locate the change, we cross initial and fine-tuned prefix sources with both continuation models, comparing one stage while holding the other fixed. Our models are adapted with LoRA on 500 GSM8K training examples. On 50 generated two-variable linear systems, three Llama runs outperform the initial checkpoint at 192 tokens but underperform it at 768; the larger loss follows the prefix source. On all 1,319 GSM8K test questions, the tested Llama run again shows a prefix-driven loss, whereas three Qwen runs improve mainly through continuation. We then remove terminal end-of-sequence (EOS) targets during training. Exchanging the learned stopping probabilities and conditional content distributions assigns 74.7–81.1% of the algebra ablation gain to stopping. EOS ablation also improves MATH-500 and HumanEval accuracy, although recovery is incomplete in some settings. Checkpoint comparisons should therefore match the intended output budget; crossed continuations help identify which stage changed.

1 INTRODUCTION

Fine-tuning gains need not persist as the output budget increases. In our algebra experiments, 100 SFT steps improve Llama-3.1-8B-Instruct by 11–13 percentage points at 192 tokens but reduce its accuracy by 9–12 points at 768, across three training seeds. Selecting by the short-budget gain would choose the worse checkpoint for longer generation. Work on test-time compute studies how to allocate inference resources (Wang et al., 2024; Snell et al., 2025); this reversal raises a complementary question: what has training changed when the apparent benefit depends on the evaluation budget?

An endpoint score cannot tell whether fine-tuning changed the partial solutions a model generates, its ability to complete a given partial solution, or both. We cross the two checkpoints' prefix sources and continuation models, keeping one stage fixed while changing the other (Figure 1a). The same pair of endpoint scores can arise from a change confined to either stage. Two constructed success matrices illustrate this ambiguity; the off-diagonal comparisons distinguish the stages under a common handoff procedure.

On algebra, Llama's reversal follows the prefix source: adapted prefixes contain more early correct answers but yield fewer final successes under either continuation model. The loss of later recoveries exceeds the gain in early successes, ruling out a shift in answer timing alone. The full GSM8K test set gives the same prefix-driven loss for Llama. Qwen, however, improves mainly through continuation. Which stage has the larger effect depends on the model and task. Additional R1-Distill-Qwen and Phi-3-mini comparisons do not reproduce the consistent Llama reversal (Appendix E).

We then test whether changes in stopping contribute to this loss of prefix value. EOS-free training has been studied under oracle length control (Newman et al., 2020), and post-SFT stopping probabilities have been adjusted to improve the accuracy–length trade-off (Zhang et al., 2025). Our matched training ablation removes terminal EOS targets; a second exchange separates the resulting stopping and conditional-content changes. The stopping gate accounts for 74.7–81.1% of the algebra EOS-ablation recovery within this policy exchange. A separate MBPP comparison tests how the policy factors affect prefix value under a fixed continuation model. On algebra, local gradient and one-step interventions connect terminal supervision to the value of continuing. EOS ablation also improves

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.

MATH-500 and HumanEval accuracy, testing the endpoint benefit beyond the small algebra set; recovery differs between the benchmarks.

2 Separating Prefix and Continuation Effects

Models, adaptation, and questions. We study Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Qwen et al., 2025), each adapted with LoRA (Hu et al., 2022) on 500 GSM8K training examples (Cobbe et al., 2021). Three 100-step runs (d100) per family provide the original budget comparison; the Llama runs also provide the algebra prefix exchange. The matched-dose algebra EOS comparison uses three training seeds at 200 steps (d200) per family. A separate d50 comparison trains standard and EOS-ablated checkpoints for 50 steps under each of three seeds per family; MATH-500 and HumanEval evaluate these same checkpoints without further training. Prompts use each model's chat template.

We use fifty fixed two-variable linear systems for exact scoring and to track recovery during continuation (16 responses per question, 800 per condition). The full 1,319-question GSM8K test split tests stage effects on natural-language mathematics (four responses per question, 5,276 per cell). A 50-question subset of this same split, with 16 responses per question, supports detailed checks of answer extraction, answer retention and recovery, first-error correction, and minimum-length decoding. The primary GSM8K stage comparisons use the full test set; the detailed scoring and intervention analyses apply to the subset.

EOS comparisons test transfer to 250 sanitized MBPP tasks (4,000 responses per condition), all 500 MATH-500 problems (Lightman et al., 2024) (eight responses each, 4,000 per condition), and all 164 HumanEval tasks (Chen et al., 2021) (eight responses each, 1,312 per condition). Unless explicitly labeled Maj@8, accuracy averages response correctness and estimates single-sample success.

Intervention and scoring. EOS-ablated SFT removes the terminal EOS target while preserving the corpus, LoRA configuration, optimizer, dose, and seed schedule. Standard and EOS-ablated d200 checkpoints are compared on algebra; the MBPP tests use Llama d100 and Qwen d200. Section 6 defines the four hybrid policies. Algebra uses strict answer extraction. For MBPP (Austin et al., 2021), our evaluator parses the full response and runs unit tests in Python 3.10 with a three-second wall limit. GSM8K uses numerical answer extraction without requiring a format marker absent from the prompt. Its exchange uses Llama d100 seed 1 and three Qwen d100 seeds on held-out test questions. Unparseable responses count as incorrect throughout. Appendix A gives the full protocols.

Notation. In C_t^L, t indexes training and L the output limit. In C_{ij}, i indexes the prefix source and j the continuation checkpoint. U and R count retained and recovered successes (Equation 5), and W is the fraction of responses that have already ended incorrectly. Subscripts 0 and t denote the initial and adapted checkpoints throughout.

Defining the gain at a given budget. Let C_t^L be the fraction of sampled responses judged correct after t training steps with output limit L. The gain over the initial checkpoint is

$$\Delta_t^L = C_t^L - C_0^L. \tag{1}$$

Within each comparison, we hold fixed the task, questions, prompt rendering, sampling settings, and extractor. We use 768 tokens as the longer tested budget, without assuming that performance has saturated there.

A positive short-budget gain reverses if the initial checkpoint benefits enough from the additional budget to overtake the adapted checkpoint. The required improvement difference is larger than the original short-budget gain. This condition describes when reversal occurs, but not which generation stage changed.

Budget and prefix comparisons. We score 192-, 384-, and 768-token prefixes of the same native long responses and separately evaluate independent 192-token generations (Table 2). In the exchange, each checkpoint supplies a prefix of up to 192 tokens and either checkpoint supplies up to 576 more. Each stored prompt and prefix token sequence is reused unchanged under both continuation checkpoints, without an inserted instruction; terminated prefixes remain in the population. All stage contrasts use the exchange's own diagonal scores (Appendix B).

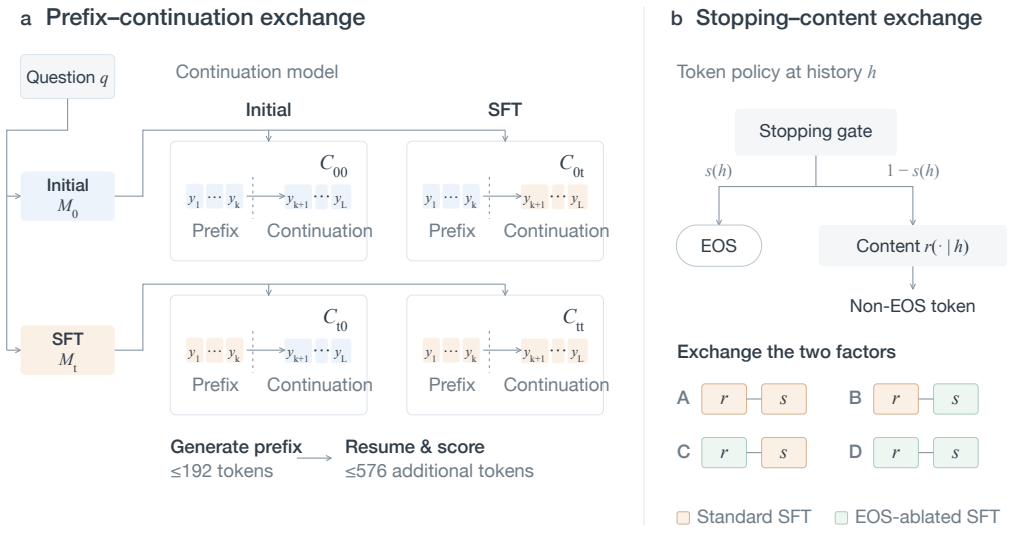

Figure 1: **Two controlled exchanges separate generation stages and policy factors.** (a) Each stored prefix is reused by both continuation models. Rows index prefix sources and columns continuation models; C_{ij} scores the complete response. Token sequences are schematic: k is the prefix length and L the output limit; generation may end earlier, and ended prefixes remain in the evaluation. (b) At history h, stopping has probability $s(h)$; otherwise the next token follows the conditional non-EOS distribution $r(\cdot \mid h)$. Policies A–D exchange these factors between standard and EOS-ablated SFT (Section 6); both factors come from checkpoint token probabilities.

Generated histories and conditional success. Let $\mu_i(h \mid q)$ be the distribution of histories handed off by checkpoint i for question q, and let $v_j(q, h)$ be the probability that checkpoint j completes that history into a response judged correct. A history includes the prompt, generated prefix, and termination state; v_j uses the actual two-stage procedure, including its handling of ended prefixes. With questions drawn uniformly from the fixed evaluation set, expected success is

$$\mathcal{C}_{ij} = \mathbb{E}_q \mathbb{E}_{h \sim \mu_i(\cdot|q)}[v_j(q, h)]. \tag{2}$$

The reported C_{ij} values estimate these expectations. Rows index prefix sources and columns continuation checkpoints, both ordered $(0, t)$. Changing one index while fixing the other gives the corresponding controlled contrast. Sequential prediction is known to depend on model-generated inputs (Ross et al., 2011); Equation 2 makes that dependence explicit for our comparison. The same non-identification result and four contrasts apply to any bounded response utility if v_j is defined as expected utility. The early-termination bound below applies to binary correctness when ended responses cannot change.

Once a history h is fixed, its source label does not enter $v_j(q, h)$. A negative source contrast therefore means that the adapted checkpoint produces a lower-value distribution of histories for the fixed continuation model.

Why endpoint scores cannot identify the changed stage. At a fixed final budget, the initial-to-initial and adapted-to-adapted scores are the two diagonal entries of the crossed matrix. They do not identify which stage changed. For any $p \neq r$ in $[0, 1]$, the two constructed success matrices

$$\underbrace{\begin{pmatrix} p & p \\ r & r \end{pmatrix}}_{\text{prefix effect only}} \quad \text{and} \quad \underbrace{\begin{pmatrix} p & r \\ p & r \end{pmatrix}}_{\text{continuation effect only}} \tag{3}$$

have the same diagonal (p, r) and training gain $r - p$. Their off-diagonal entries distinguish the two explanations. Appendix G constructs both cases.

Comparing the two stages. The crossed matrix directly reports two source contrasts, $C_{tj} - C_{0j}$, and two continuation contrasts, $C_{it} - C_{i0}$. Its diagonal difference has two valid paths,

$$C_{tt} - C_{00} = (C_{t0} - C_{00}) + (C_{tt} - C_{t0}) = (C_{0t} - C_{00}) + (C_{tt} - C_{0t}). \tag{4}$$

When the interaction $J = C_{tt} - C_{t0} - C_{0t} + C_{00}$ is nonzero, the two paths allocate different amounts to the stages. We therefore report all four controlled contrasts. The design requires a sequential generator that can resume a stored history and whose completed responses can be scored. It assumes neither monotonicity nor a parametric form for continuation quality.

Early stopping and checkpoint selection. Let W_m^L be the fraction of responses that have ended incorrectly by token L. For native generation with absorbing termination and an unchanged decoding policy, later accuracy satisfies $C_m^H \leq 1 - W_m^L$ for $H \geq L$. Thus, for the same response population, $1 - W_t^L < C_0^H$ rules out the adapted checkpoint at budget H. Appendix K gives the assumptions and application to the recorded trajectories.

3 Short-Budget Gains and Fewer Later Recoveries

All three adapted Llama checkpoints outperform the initial model by 11.000–13.125 points at 192 tokens and underperform it by 9.000–12.000 points at 768 (Appendix Table 2). Selecting an adapter by the sign of its gain would give opposite decisions at the two budgets. Qwen's losses shrink from 6.750–8.875 points to 0.625–1.375 points at 768; the larger budget changes the magnitude without changing the preferred checkpoint. Its initial 768-token accuracy is already 99.375%, leaving little headroom for improvement. At the intermediate 384-token cutoff, Llama already has a negative gain, while Qwen has a larger deficit than at either endpoint (Appendix Table 3).

Where the additional successes come from. For Llama seed 1, truncating the native long responses gives 181 correct initial-model responses at 192 tokens and 599 at 768. The adapted checkpoint moves from 275 to 513. In both states, no correct response is lost between these two inspected budgets. The initial model therefore recovers 418 answers, while the adapted model recovers 238. The budget-gain difference is $(238 - 418)/800 = -22.5$ points, changing the paired short-budget gain from $+11.75$ to -10.75 points.

Qwen seed 1 gains more recoveries after adaptation. Its initial model recovers 601 responses, compared with 644 for the adapted checkpoint. Those 43 additional recoveries reduce the loss by 5.375 points. A short cutoff can therefore conceal either an increase or a decrease in later recoveries.

Correct answers first appear earlier after Llama adaptation (median token 227 to 173) and later after Qwen adaptation (252 to 298). These medians measure the time of the first extractable correct answer, conditional on success within the tested budget; both answer content and presentation can affect them. The exchange below tests the value of the entire prefix distribution under the same scorer.

4 Locating the Loss Before Continuation

Holding the continuation model fixed tests whether the recovery deficit follows the prefix source.

Retained and recovered answers. Let P_i be the number of correct prefixes from source i, U_{ij} the number of those successes retained after checkpoint j continues them, and R_{ij} the number of initially incorrect prefixes that become correct. For $N = 800$ responses,

$$N(C_{tj} - C_{0j}) = (U_{tj} - U_{0j}) + (R_{tj} - R_{0j}). \tag{5}$$

The first term compares retained prefix successes; the second compares recoveries from initially incorrect prefixes. Dividing R_{ij} by the number of incorrect prefixes gives the recovery rate for the source population under continuation checkpoint j.

For seed 1, initial prefixes attain 75.750% success when the initial model continues them and 72.125% when the adapted model continues them. Adapted prefixes attain 65.500% and 64.375%, respectively. Switching only the prefix source therefore costs 10.250 points with the initial continuation model, or 7.750 points with the adapted continuation model. Switching only the continuation model costs

3.625 points on initial prefixes and 1.125 on adapted prefixes. Seeds 2 and 3 give the same ordering of effect sizes; Appendix B gives exact counts.

In each seed, the smaller absolute prefix-source contrast exceeds the larger absolute continuation contrast by 4.125, 4.750, and 6.750 points. Across these three runs, the larger measured loss follows the prefix source under the tested handoff procedure.

The interaction J is $+2.500$, $+1.875$, and $+3.500$ points across the algebra seeds, meaning that the adapted continuation model loses less on adapted-source histories. Averaged across seeds, the smaller absolute source effect exceeds the larger continuation effect by 5.208 points. In paired question resampling, the 95% interval for this ordering margin is $[0.42, 9.71]$ (Appendix B); the estimate is conditional on the evaluated checkpoints and handoff procedure.

Table 18 and Appendix Figure 4 account for the algebra reversal. Adapted Llama prefixes add 88–105 correct answers by token 192, and every one remains correct after continuation. Yet under the initial continuation model they produce 175–195 fewer recoveries, leaving 78–107 fewer final answers despite the early lead. Under the adapted continuation model they produce 155–168 fewer recoveries and 62–79 fewer final answers. Among prefix-incorrect histories, recovery falls from 68.7% for initial prefixes to 43.3–47.5% for adapted prefixes under the initial continuation model. With the adapted continuation model, the corresponding ranges are 64.0–64.6% and 43.9–45.8%.

A lower conditional recovery rate could reflect depletion: easy responses may leave the unresolved set earlier. The count identity tests whether earlier arrival of the same successful answers can account for the final contrast. If the 88–105 additional early successes were only existing 768-token successes appearing sooner, then $\Delta R = -\Delta U$ and $\Delta C = 0$. Instead, the recovery loss exceeds the entire retained-success gain by 78–107 answers under the initial continuation model and by 62–79 under the adapted model. The adapted prefixes therefore have lower final value under either continuation model despite containing more early successes. Differences in the composition of their unresolved histories can still contribute to the conditional recovery rates.

The loss also appears among still-active seed 1 prefixes. At token 192 in the exchange collection, the initial checkpoint finishes correctly on 427/586 (72.9%) initial-source histories but only 249/455 (54.7%) adapted-source histories. With the adapted checkpoint continuing both sets, the corresponding rates are 398/586 (67.9%) and 241/455 (53.0%). The active populations depend on the prefix source, so these rates compare different sets of histories.

A correctable error need not explain the model gap. To test whether correcting intermediate content closes the gap, we replace the first verifiable error in eligible prefixes from the initial and standard-SFT d100 seed-1 checkpoints. Relative to a generic check prompt, correction adds 91 successful algebra continuations for each checkpoint when it continues its own eligible prefixes. Holding unselected outputs unchanged therefore gives zero change in the constructed 800-response model gap; on the 50-question GSM8K diagnostic subset the change is -0.375 points. Thus correcting the selected errors improves both checkpoints without closing this constructed adaptation gap. Appendix H specifies the treatments, and Appendix I derives the comparison. These experiments use standard-SFT checkpoints and do not explain the residual loss after EOS ablation.

5 Different Stage Effects on Held-Out GSM8K

The algebra comparison locates a prefix-driven loss, but does not establish that the same stage dominates on other tasks. We test this on the full GSM8K test set using Llama d100 seed 1 and all three Qwen d100 seeds, with the same 192+576-token procedure and 5,276 responses per cell. The full-response scoring rule extracts an explicit answer when present and otherwise the final number. In the original 50-question subset, the prior predictions of a short-budget Llama gain and a shrinking Qwen loss did not hold; the predicted difference in stage ordering did (Appendix C.1).

Llama's larger loss occurs before continuation. Changing from initial to adapted prefixes costs 7.127 points with the initial continuation model and 6.710 points with the adapted continuation model. Changing the continuation model instead gives -0.208 points on initial prefixes and $+0.208$ on adapted prefixes. The smaller prefix effect exceeds the larger continuation effect in absolute value by

6.501 points (95% percentile bootstrap interval $[5.36, 7.34]$). Thus Llama's larger loss follows the prefix source on the full test set, as on algebra.

Qwen benefits mainly from continuation. Across the three Qwen seeds, changing to the adapted continuation model adds 4.587–5.099 points on initial prefixes and 4.321–4.587 on adapted prefixes. Changing the prefix source adds only 0.284–0.929 points. The smaller continuation effect exceeds the larger prefix effect by 3.393, 3.848, and 3.791 points across seeds; the gains between the two diagonal conditions are 5.250, 5.023, and 5.383 points. The 95% percentile bootstrap intervals for these margins are $[2.41, 4.08]$, $[2.91, 4.53]$, and $[2.82, 4.47]$, respectively, supporting the aggregate continuation ordering in each seed. This ordering describes average effects, not every question; the original subset shows heterogeneous question-level contrasts (Appendix C.2).

Across all seven algebra and GSM8K comparisons, continuing initial prefixes with the adapted checkpoint outperforms the reverse combination, but never exceeds the best unmixed checkpoint. Appendix C gives the full-test counts and intervals. On the 50-question diagnostic subset, Llama's loss persists with final-number extraction but reverses when extraction requires an answer marker or cue, illustrating the dependence on scoring rules (Appendix C.1).

6 Separating Stopping and Content Changes

For Llama, the stage comparisons locate the loss in generated histories. They leave open whether content, stopping, or both account for the change. On algebra, 217/800 initial Llama responses end by token 192, compared with 346–349 after d100 adaptation. Qwen moves from 195/800 to 120–138, with a smaller loss at the longer budget. We test the stopping contribution using a matched training ablation and then exchange the two factors of the learned token policy.

For a target answer with T tokens, terminal supervision includes

$$\mathcal{L}_{\text{term}}(\theta) = -\log q_\theta(\text{EOS} \mid x, y_{1:T}). \tag{6}$$

The fixed 500-example corpus has median target lengths of 68 tokens for Llama and 79 for Qwen; 494 and 488 targets, respectively, end within token 192. The terminal loss rewards ending after these completed demonstrations. Evaluation instead encounters generated histories that may still be unresolved at the same length. We test this possible mismatch in the matched d200 comparison, where both families have a long-budget deficit.

Table 1: **Removing terminal supervision mitigates long-budget losses and raises prefix value under a fixed initial continuation model.** Endpoint columns compare each adapted checkpoint with its initial checkpoint; the source column subtracts standard-SFT prefix value from EOS-ablated prefix value under the unchanged initial continuation model. All entries are means over three training seeds, in percentage points; dashes mark comparisons not reported here. Appendix A.1 gives the per-seed results.

Model/steps	Task	Std. ΔC^{768}	$-$EOS ΔC^{768}	Endpoint Δ	Source Δ
Llama/d200	Algebra	-36.667	-15.000	$+21.667$	—
Qwen/d200	Algebra	-27.667	-9.500	$+18.167$	—
Llama/d100	Algebra	-10.583	-3.458	$+7.125$	—
Llama/d100	MBPP	-6.142	-2.408	$+3.733$	$+3.650$
Qwen/d200	MBPP	—	—	—	$+6.317$

Removing the terminal target reduces the d200 long-budget loss in both families (Table 1). To distinguish the resulting changes in stopping and content, we exchange two factors of the learned token distribution throughout the 768-token algebra generation. For a checkpoint trained under condition z, write its EOS probability as $s_z(h)$ and, when $s_z(h) < 1$, its conditional distribution over non-EOS tokens as $r_z(a \mid h)$. The hybrid policy with content source i and stopping gate j is

$$q_{ij}(a \mid h) = \begin{cases} s_j(h), & a = \text{EOS}, \\ [1 - s_j(h)] r_i(a \mid h), & a \neq \text{EOS}. \end{cases} \tag{7}$$

Let A, B, C, D denote $(r_{\text{std}}, s_{\text{std}})$, $(r_{\text{std}}, s_{-\text{EOS}})$, $(r_{-\text{EOS}}, s_{\text{std}})$, and $(r_{-\text{EOS}}, s_{-\text{EOS}})$. For their success rates, the symmetric stopping-gate and content allocations are

$$G = \tfrac{1}{2}[(B - A) + (D - C)], \qquad K = \tfrac{1}{2}[(C - A) + (D - B)], \tag{8}$$

so $G + K = D - A$. We average the two replacement orders and report all four controlled contrasts. The denominator $D - A$ is the recovery from removing EOS targets, not the full fine-tuning deficit relative to the initial checkpoint. The initial-model score does not enter either allocation or its denominator.

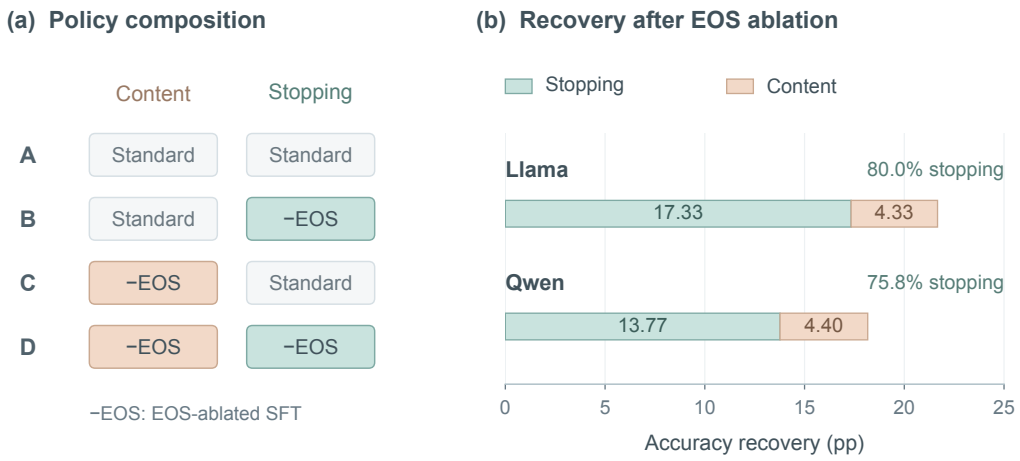

Figure 2: **Stopping contributes most of the EOS-ablation recovery.** (a) Four policies combine conditional non-EOS content distributions and EOS stopping probabilities from standard and EOS-ablated SFT checkpoints. (b) Mean allocations G and K from Equation 8 across three d200 algebra seeds, with 800 responses per cell and seed at 768 tokens. Stopping share is $\overline{G}/(\overline{G} + \overline{K})$, relative to standard SFT. Appendix Table 6 reports every seed's four counts.

Across the six seed-level comparisons summarized in Figure 2, G ranges from 13.625 to 17.750 points and accounts for 74.7–81.1% of $D - A$. Every controlled gate effect exceeds either controlled content effect. Changing only the gate increases mean length by 39.5–42.7 tokens; changing only content changes it by 0.9–3.3 tokens. These contrasts measure changes in generation length and accuracy, not parameter or distributional distance.

On MBPP, all four hybrid policies generate prefixes of up to 192 tokens; the unchanged initial checkpoint supplies up to 576 more. For a fixed family and training dose, let t_z denote the checkpoint trained under $z \in \{\mathrm{std}, -\mathrm{EOS}\}$. Define $\Gamma_0^z = \mathcal{C}_{t_z 0} - \mathcal{C}_{00}$. Consequently,

$$\Gamma_0^{-\mathrm{EOS}} - \Gamma_0^{\mathrm{std}} = \mathbb{E}_{\mu_{t_{-\mathrm{EOS}}}} v_0 - \mathbb{E}_{\mu_{t_{\mathrm{std}}}} v_0, \tag{9}$$

with the common question average implicit. Across the three seeds, EOS ablation raises MBPP prefix value by 3.55–3.75 points for Llama and 6.15–6.45 for Qwen, under the unchanged initial continuation checkpoint. Appendix Tables 9 and 10 give the counts.

The gate–content exchange on MBPP gives the same ordering as on algebra. Gate allocations are 2.738–2.888 points for Llama d100 and 4.900–5.125 for Qwen d200, versus content allocations of 0.813–0.863 and 1.250–1.325. Every controlled gate contrast exceeds either content contrast. Gate shares range from 76.7% to 79.7% (Appendix Table 11). These are effects on prefix value under the fixed handoff and full-text scorer, including any interaction between a completed program prefix and appended code. The aggregate comparison does not isolate that interaction from changes in unfinished programs.

In these algebra and MBPP comparisons, EOS ablation improves every seed, but source contrasts and adapted native endpoints remain below their initial-checkpoint controls (Table 1; Appendix A.1). Across all twelve gate–content comparisons, deleting any one question cannot reverse the EOS benefit or the gate-over-content ordering (Appendix M).

When continuing helps. Whether a gate change helps depends on the value of continuing. If $u(h)$ scores ending the controlled segment and $Q(h)$ its reference continuation value, the local effect is $\delta s(h)[u(h) - Q(h)]$. Appendix L derives the exact trajectory sum. Lowering EOS probability helps

where continuation is more valuable and can hurt where it spoils a correct answer. The four-cell contrasts measure this net effect, rather than length alone.

We separate the EOS and content losses: terminal targets reward stopping, while content targets reward continuing and token identity. At each family's standard d200 seed-1 checkpoint, we measure their local gradient effects on EOS log-odds in two groups of 20 algebra prefixes. The EOS term pushes toward stopping on 19/20 and 18/20 Llama prefixes and 18/20 and 17/20 Qwen prefixes; the content term has negative mean effects throughout. On the same histories, EOS ablation also lowers mean stopping probability in all four groups. On prefixes preceding incorrect early termination, suppressing EOS for one step raises continuation success by 6.88 points in Llama and 5.63 in Qwen; prefixes preceding correct termination lose 2.81 and 1.25 points. The intervention recovers some incorrect answers but can spoil correct ones (Appendix L).

A decoding intervention. With the Llama d100 seed-1 parameters fixed, suppressing EOS before token 192 reduces the model deficit from 86 to 33 responses on algebra and from 81 to 36 on the 50-question GSM8K diagnostic subset (reductions of 61.6% and 55.6% relative to the respective original deficits). Both checkpoints receive the intervention. This changes later histories as well as length, so its effect cannot be added to the training ablation as a mediation fraction. Appendix F gives the policy and controls.

Transfer and residual loss. On MATH-500 at d50, EOS ablation improves response accuracy in all three seeds of both families, by 3.57 points on average for Llama and 1.23 for Qwen. These mean gains close 12.94% and 6.19% of the initial-to-SFT deficits. Maj@8 gains are positive in every seed; for Qwen, the seed-2 question-bootstrap interval reaches zero and the seed-3 interval crosses it (Appendix A.2).

On HumanEval, using the same d50 checkpoints as MATH-500 without further training, EOS ablation improves on standard SFT by 6.68 and 6.07 points on average for Llama and Qwen. Every ablated checkpoint also exceeds its initial reference (Appendix A.3). These benchmark results extend the endpoint benefit, not the gate–content allocation. Because training doses and decoding settings differ across tasks, differences in gain magnitude cannot be attributed to the task alone.

7 Scope and Limitations

The stage contrasts are conditional on the tested continuation models and handoff procedure. A greedy MATH-500 diagnostic finds small mean excess negative log-likelihood (NLL) changes near the handoff among window-complete pairs (Appendix A.4); it does not establish the absence of distribution shift under sampled decoding. Different tokenizers or a history-dependent continuation policy define another comparison. The framework allows lost prefix successes, but does not measure evaluator-independent reasoning ability.

The framework applies to resumable sequential generation with a fixed bounded response utility, including non-binary utilities in the gate identity. The main experiments cover two 7B/8B families after short LoRA adaptation on 500 GSM8K examples, not full-scale post-training recipes. Evaluation spans 50 generated two-variable linear-system questions, the 1,319-question GSM8K test split, 500 MATH-500 problems, 250 MBPP tasks, and 164 HumanEval tasks. Repeated sampling estimates decoding variability on these questions without increasing the diversity of algebra problems.

The MATH-500 and HumanEval t intervals summarize three paired training seeds on fixed benchmarks, rather than uncertainty over new questions. Full-test GSM8K covers one Llama training seed and three Qwen seeds, so it does not estimate Llama's across-seed variability. Its different stage orderings show that prefix effects need not dominate; scores near zero or one can also obscure the contrasts.

The gate–content exchanges cover three seeds on algebra and MBPP; the gradient and one-step interventions use d200 seed 1, with two groups of 20 incorrect algebra prefixes and 20 correct-end controls per family. The gradients measure local SGD sensitivity, not the accumulated optimizer trajectory. Neither result establishes gate dominance under other objectives, architectures, scales, or decoding rules.

Among the six d100 algebra comparisons on the common recorded question set, selecting by the short-budget gain favors all three adapted Llama checkpoints, which underperform at 768 tokens; both budgets favor initial Qwen in its three comparisons. The early-termination bound is inconclusive for the three Llama d100 checkpoints. These records therefore do not demonstrate a successful early rejection of an inferior checkpoint (Appendix K).

8 Related Work

Sequential prediction depends on earlier model-generated inputs (Ross et al., 2011; Agarwal et al., 2024). Exposure-bias analyses compare continuation from ground-truth and generated prefixes (He et al., 2021), while prefix distillation supervises prefixes of student-generated trajectories (Zhang et al., 2026). Crossing the initial and adapted checkpoints separates effects of prefix generation from those of continuation.

Compute-matched evaluation can change comparisons between reasoning strategies (Wang et al., 2024), and brevity constraints can reverse model hierarchies (Hakim, 2026). Test-time scaling studies how to allocate inference resources (Snell et al., 2025); s1 controls the reasoning-token budget through budget forcing (Muennighoff et al., 2025), while work on overthinking identifies redundant reasoning (Chen et al., 2025). We examine a within-model fine-tuning gain at a fixed sample count: paired responses track the gain across output budgets, and the prefix exchange compares the two stage effects.

Newman et al. (2020) compare EOS and EOS-free training under oracle length control and analyze hidden-state dynamics. Zhang et al. (2025) adjust post-SFT stopping probabilities and use length-regularized reinforcement learning. Our policy exchange distinguishes the controlled effects of learned stopping probabilities and conditional content distributions on algebra endpoints and MBPP prefix value, without inferring the same hidden-state mechanism.

SFT gains can reflect format learning (Yang et al., 2025); evaluation also depends on answer extraction (Jo et al., 2025) and output constraints (Deng et al., 2026). Our local-edit experiment tests whether replacing an intermediate error closes the adaptation gap. It relates to error backtracking (Tyen et al., 2024; Zhu et al., 2026), first-error replacement (Yang et al., 2026), and limits of SFT-based self-correction (Kumar et al., 2025). Appendix N discusses further connections to adaptation, decoding, and evaluation.

9 Conclusion

Fine-tuning gains depend on the output budget and can reverse as generation continues. The crossed comparisons locate the larger change: on algebra, Llama loses more later recoveries than it gains early successes under either continuation model. The tested Llama run shows a prefix-driven loss on GSM8K as well, whereas Qwen gains mainly through continuation in all three training runs.

Stopping is one contributor to these changes. Removing terminal EOS targets improves long-budget outcomes, and the algebra policy exchange assigns more of this improvement to stopping than to conditional content. Local interventions show that continuing can recover an error or spoil a correct answer. These findings favor checkpoint selection at the intended output budget, with stage-specific comparisons guiding where to intervene.

Reproducibility statement

The appendix specifies the training, decoding, scoring, and prefix-handoff protocols and reports the counts underlying the experimental comparisons. Question-level bootstrap procedures preserve paired conditions and repeated responses within each question; the EOS gate–content allocations are reported as seed-level point estimates. MATH-500 and HumanEval response-accuracy intervals summarize three paired training-seed differences, while MATH-500 voting intervals use paired questions. The corresponding sections give the resampling procedures and statistical units.

ETHICS STATEMENT

The study evaluates public model checkpoints on generated task answers. Executable-code evaluation requires isolation and resource limits. Task scores under the tested decoding conditions do not certify safe deployment.

AI USE STATEMENT

AI assistance was used for feedback on hypotheses and experimental design, experiment execution and analysis, interpretation of results, literature verification, inspection of research records, mathematical derivations and proof checking, numerical checks, manuscript revision (including LLM assistance from GPT-5.6 with grammar and wording), and plotting and document-building code. The authors are responsible for the experimental records, scientific claims, and final manuscript.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219v4. Version 4, August 2024.

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 *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=aloEru2qCG.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, et al. Lessons from the trenches on reproducible evaluation of language models, 2024b. URL https://arxiv.org/abs/2405.14782v2. Version 2, May 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374v2.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 9487–9499, 2025. URL https://proceedings.mlr.press/v267/chen25bx.html.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL https://www.jmlr.org/papers/v25/23-0870.html.

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. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Haikang Deng, Po-Nien Kung, and Nanyun Peng. Decoupling task-solving and output formatting in LLM generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16764–16781, 2026. doi: 10.18653/v1/2026.acl-long.764. URL https://aclanthology.org/2026.acl-long.764/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783v3. Version 3, November 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

MD Azizul Hakim. Brevity constraints reverse performance hierarchies in language models. *arXiv preprint arXiv:2604.00025*, 2026. URL https://arxiv.org/abs/2604.00025.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5087–5102, 2021. doi: 10.18653/v1/2021.emnlp-main.415. URL https://aclanthology.org/2021.emnlp-main.415/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5254–5276. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.319. URL https://aclanthology.org/2023.emnlp-main.319/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations*, pp. 32808–32824, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Hwiyeol Jo, Joosung Lee, Jaehong Lee, Sang-Woo Lee, Joonsuk Park, and Kang Min Yoo. Finding answers in thought matters: Revisiting evaluation on large language models with reasoning. *arXiv preprint arXiv:2510.14773*, 2025. URL https://arxiv.org/abs/2510.14773.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/871ac99fdc5282d0301934d23945ebaa-Abstract-Conference.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW.

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 *International Conference on Learning Representations*, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The Flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2173–2185, 2020. doi: 10.18653/v1/2020.emnlp-main.170. URL https://aclanthology.org/2020.emnlp-main.170/.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 276–291, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.26. URL https://aclanthology.org/2020.blackboxnlp-1.26/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115v2. Version 2, January 2025; first posted December 2024.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06732v7.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, volume 38, pp. 174627–174662, 2025. doi: 10.52202/085713-5808. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/ff541950d1e885af90f523571564a401-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *International Conference on Learning Representations*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Felix Stahlberg and Bill Byrne. On NMT Search Errors and Model Errors: Cat Got Your Tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, 2019. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331/.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.826. URL https://aclanthology.org/2024.findings-acl.826/.

Junlin Wang, Siddhartha Jain, Ben Athiwaratkun, Dejiao Zhang, Baishakhi Ray, and Varun Kumar. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19916–19939, 2024. doi: 10.18653/v1/2024.emnlp-main.1112. URL https://aclanthology.org/2024.emnlp-main.1112/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

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 *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. Consistency of a Recurrent Language Model With Respect to Incomplete Decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5553–5568, 2020. doi: 10.18653/v1/2020.emnlp-main.448. URL https://aclanthology.org/2020.emnlp-main.448/.

Linyao Yang, Jian-Tao Huang, Yafei Lu, Zhenhui Jessie Li, and Guirong Xue. The emperor's new reasoning: Format imitation overshadows genuine mathematical understanding in SFT. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21087–21100, 2025. doi: 10.18653/v1/2025.emnlp-main.1067. URL https://aclanthology.org/2025.emnlp-main.1067/.

Matthew Y. R. Yang, Hao Bai, Ian Wu, Gene Yang, Amrith Setlur, and Aviral Kumar. InT: Self-proposed interventions enable credit assignment in LLM reasoning. In *International Conference on Learning Representations*, pp. 85054–85091, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/89062e4d480c0c3a88d36c20c5694459-Abstract-Conference.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy

distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026. doi: 10.18653/v1/2026.findings-acl.1276. URL https://aclanthology.org/2026.findings-acl.1276/.

Xuechen Zhang, Zijian Huang, Chenshun Ni, Ziyang Xiong, Jiasi Chen, and Samet Oymak. Making small language models efficient reasoners: Intervention, supervision, reinforcement. *arXiv preprint arXiv:2505.07961*, 2025. URL https://arxiv.org/abs/2505.07961v3.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55006–55021, 2023. doi: 10.52202/075280-2400. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. Dissecting failure dynamics in large language model reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8893–8914, 2026. doi: 10.18653/v1/2026.acl-long.401. URL https://aclanthology.org/2026.acl-long.401/.

A METHODS AND ANALYSIS SCOPE

Adaptation and decoding. SFT uses 500 GSM8K training examples with calculator annotations removed, corpus seed 20260828, LoRA rank 16, scaling 32, dropout 0.05, learning rate 10^{-4}, microbatch size four, gradient accumulation factor two, maximum sequence length 1024, and bfloat16 arithmetic. The main Llama and Qwen SFT runs use AdamW with linear learning-rate decay, no warmup, and zero weight decay; these settings are identical within each standard/EOS-ablated pair. For both Llama and Qwen, LoRA targets `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, with `bias=none` and `use_rslora=False`. Training checks that LoRA is attached to every requested module type; the saved checkpoint configurations confirm the settings.

The d50, d100, and d200 runs start from their respective initial checkpoints. For d50, standard and EOS-ablated SFT each use three training seeds per family, with the same GSM8K-500 corpus and adaptation settings; their merged checkpoints are reused for both MATH-500 and HumanEval. The original prefix-exchange analysis uses the 100-step Llama runs. The EOS analysis includes a matched d200 algebra comparison in both families, additional Llama d100 endpoint results, and MBPP source comparisons for Llama d100 and Qwen d200.

For the target-length analysis around Equation 6, we reproduce the same corpus selection from the official GSM8K training split and remove calculator annotations before applying each model's chat template. The Llama tokenizer gives a mean target length of 75.634 tokens, median 68, 10th/25th/75th/90th percentiles of 37/48/94/126, and maximum 242; 494/500 targets are at most 192 tokens and 463 are at most 138. The Qwen tokenizer gives mean 88.4, median 79, percentiles 44/57/108/145, and maximum 301; the corresponding counts are 488 and 442.

The algebra items are programmatically generated with seed 42, not drawn from a public benchmark. The primary set P contains 50 two-variable linear-system questions, with 16 sampled responses per question. Scoring one response at several budgets or with several extractors does not add independent questions. The paired-budget and elicitation analyses use eight checkpoints: two initial checkpoints and three adaptations per model. The Llama prefix exchange adds ten continuation conditions, with initial-to-initial generation shared across seeds.

The main generation records use chat serialization, temperature 0.8, top-$p = 1$, and generation sampling seed 42. For the answer-elicitation experiment, the branch metadata specifies a nominal prefix length of 192, a 64-token extension allowance, and replay generation to 256 tokens. This differs from the 192+576-token prefix exchange. Answer-now and instructed continuation append different textual instructions; bare continuation is obtained by replaying the native stream. The branches have the same output cap but can differ in instruction length and actual token use.

Scoring and uncertainty. For the two-variable algebra task, strict scoring finds the last case-insensitive match of `answer\s*:` and reads only the first line following it, after trimming surrounding whitespace and asterisks. It extracts variable–integer assignments with the case-insensitive pattern `([xyz])\s*=\s*(-?\d+)`. Variable names are lowercased and repeated names take their last matched value. The resulting dictionary must equal the reference dictionary exactly: both x and y must have their reference values, with no additional variable. No marker or no matching assignments gives an incorrect response. For example, `ANSWER: x=2, y=-1` matches this format. The rule performs literal integer extraction, not symbolic equation solving.

Expanded scoring is a supplementary sensitivity analysis with an alternative extractor. Both rules average correctness over all sampled responses, including invalid or absent answers; neither selects an answer by searching for a match to the gold value.

The post-hoc full-response analysis uses 10,000 paired question-bootstrap draws with resampling seed 20260911. The resampling unit is the question: each draw retains its 16 outputs across states and branches, averaging adapted seeds within the draw for Equation 13. Thus the algebra bootstrap resamples 50 question clusters, not 800 independent observations. Percentile intervals describe variation over questions for the observed checkpoints, seeds, and procedure. Intervals throughout the paper cover individual contrasts, without adjustment for simultaneous inference. They do not estimate variation over model families or new training seeds. We also report the three adapted estimates separately.

Scoring consistency. The records contain budget and continuation judgments, answer transitions, and comparisons between independent short generations and truncated prefixes. The full-response branch file contains 6,400 rows. Both extractors preserve the prefix judgment in all 3,728 cases with an empty continuation, confirming that rescoring an unchanged response preserves its label. These annotations do not reconstruct the original generation and judging process or independently validate every semantic label. The first-error tables summarize the rule-based annotation and single-annotator review described in Appendix D.1.

Table 2: **The reversal appears in all three Llama training runs.** Each score uses 800 responses. Initial scores are shared across the three adapted checkpoints, rather than three independent initial-model runs. Gains are in percentage points. The C^{192} column scores independent 192-token generations; the final column instead scores the first 192 tokens of native 768-token responses. For Llama seed 1 these give 274/800 (34.250%) and 275/800 (34.375%), respectively. The exchange uses 274 prefix successes (Appendix B).

Model	State	C^{192} (%)	C^{768} (%)	$\Delta^{192}/\Delta^{768}$ (pp)	Prefix C^{192} (%)
Llama	Initial	22.625	74.875	—	22.625
	Seed 1	34.250	64.125	$+11.625/-10.750$	34.375
	Seed 2	35.750	65.875	$+13.125/-9.000$	35.750
	Seed 3	33.625	62.875	$+11.000/-12.000$	33.625
Qwen	Initial	24.250	99.375	—	24.250
	Seed 1	17.500	98.000	$-6.750/-1.375$	17.500
	Seed 2	17.250	98.375	$-7.000/-1.000$	17.250
	Seed 3	15.375	98.750	$-8.875/-0.625$	15.375

Table 3: **The intermediate budget distinguishes the two gain trajectories.** Strict correct counts from the same 800 native 768-token responses per checkpoint, truncated at each budget. All three columns use paired-prefix scoring; the 192-token column is distinct from the independent short-generation column of Table 2.

Model	State	192 tokens	384 tokens	768 tokens
Llama	Initial	181	565	599
	Seed 1	275	501	513
	Seed 2	286	514	527
	Seed 3	269	488	503
Qwen	Initial	194	756	795
	Seed 1	140	636	784
	Seed 2	138	629	787
	Seed 3	123	628	790

A.1 Training-side EOS ablation

The matched algebra comparison uses Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct at d200. Additional seed-level comparisons use Llama d100 on algebra and MBPP and Qwen d200 on MBPP. Within each comparison, standard and EOS-ablated SFT share the 500-example corpus, LoRA configuration, optimizer, step count, and seed schedule; the ablation masks terminal EOS loss while retaining the original normalization denominator. Algebra uses strict answer extraction, and MBPP uses the code-extraction and unit-test judge specified below. The d200 algebra comparison uses 50 questions and 16 responses per question. Each adapted training seed is paired with a separately sampled initial-checkpoint control on the same questions and scoring rule. Table 4 reports its seed-level counts; Tables 6 and 7 report the stopping-gate–content cross. Tables 8, 9, and 10 report the additional algebra and MBPP comparisons.

What the intervention identifies. Equation 9 holds the evaluator v_0 fixed and changes the training target used by the prefix producer. On MBPP, the resulting effect includes changes in both prefix content and termination. The d200 algebra comparison instead exchanges the learned EOS probability

and conditional non-EOS distribution at every history. The minimum-length experiment changes decoding while holding checkpoint parameters fixed; continuing past a stopping point can change the subsequent history. The local-error experiment uses the standard-SFT d100 checkpoints, not EOS-ablated checkpoints, so it does not identify the remaining post-ablation loss. These interventions answer different questions and do not form an additive partition of the full fine-tuning deficit.

Seed-level comparisons. At d200, EOS ablation reduces the algebra long-budget loss in every seed of both families (Table 4). We report endpoint counts for each seed and paired response-outcome counts for Qwen d200 seeds 2 and 3 (Table 5). For each seed, the gate–content comparison reports four correct-response counts, mean generation lengths, and descriptive allocations. Tables 8–10 give the additional d100 algebra and MBPP counts or count differences.

Table 4: **At the matched 200-step dose, EOS removal reduces the algebra long-budget loss in every seed of both families.** Each row uses 800 responses. Initial controls are sampled separately for each seed on the same 50 questions. All accuracy differences are in percentage points at 768 tokens. The two change columns compare each trained checkpoint with that row's initial control; Gain compares EOS-ablated with standard SFT.

Model	Seed	Init.	Std.	−EOS	Δ_{std}	$\Delta_{-\text{EOS}}$	Gain
Llama	1	597	301	473	−37.000	−15.500	+21.500
	2	604	312	487	−36.500	−14.625	+21.875
	3	600	308	481	−36.500	−14.875	+21.625
Qwen	1	793	571	715	−27.750	−9.750	+18.000
	2	797	576	722	−27.625	−9.375	+18.250
	3	795	574	720	−27.625	−9.375	+18.250
Llama	Mean	600.3	307.0	480.3	−36.667	−15.000	+21.667
Qwen	Mean	795.0	573.7	719.0	−27.667	−9.500	+18.167

Paired response outcomes. For Qwen d200 seeds 2 and 3, Table 5 summarizes matched response outcomes under standard and EOS-ablated SFT. Seed 2 has 152 wrong-to-correct and six correct-to-wrong pairs; seed 3 has 155 and nine. Both gain 146 correct responses through different paired transitions. Among standard-SFT failures, the fractions that become correct are $152/224 = 67.86\%$ and $155/226 = 68.58\%$. Each transition compares two sampled responses; it does not count a distinct question or an edit to the original answer. The ablated checkpoints remain below their initial controls by 75 correct responses in each of these two seeds; the seed-1 deficit is 78 (Table 4).

Table 5: **EOS ablation corrects more responses than it spoils in both Qwen seeds.** Paired Qwen d200 algebra outcomes, with 800 sampled-response pairs per seed on 50 questions. Transitions run from standard to EOS-ablated SFT. Correct/incorrect refer to final response scores. The last column divides wrong-to-correct pairs by all standard-SFT incorrect responses.

Seed	Both correct	Wrong→correct	Correct→wrong	Both wrong	Net gain	Recovery rate
2	570	152	6	72	+146	67.86%
3	565	155	9	71	+146	68.58%

Crossing the stopping gate and conditional content distribution. For the d200 algebra checkpoints, let $s_z(h) = q_z(\text{EOS} \mid h)$ and, when $s_z(h) < 1$, let $r_z(a \mid h) = q_z(a \mid h)/[1 - s_z(h)]$ for $a \neq \text{EOS}$, where z is standard or EOS-ablated SFT. Equation 7 combines the stopping gate from one checkpoint with the conditional non-EOS distribution from the other at every generated history. The four policies are

$$A = (r_{\text{std}}, s_{\text{std}}), \qquad B = (r_{\text{std}}, s_{-\text{EOS}}),$$
$$C = (r_{-\text{EOS}}, s_{\text{std}}), \qquad D = (r_{-\text{EOS}}, s_{-\text{EOS}}).$$

A and D reproduce the standard and EOS-ablated endpoint counts in Table 4; B and C change one policy factor. Every cell contains 800 responses on the same 50 questions under the 768-token cap

and strict algebra scorer. The off-diagonal policies can generate histories that are unlikely under either original checkpoint. Their contrasts include these changes in the history distribution and do not isolate a history-independent training mechanism.

Table 6: **The stopping gate contributes most of the d200 EOS-ablation gain in every seed.** A–D are correct counts out of 800. G and K are the symmetric percentage-point allocations in Equation 8; gate share divides G by the percentage-point change represented by $D - A$.

Model	Seed	A	B	C	D	G (pp)	K (pp)	Gate share
Llama	1	301	433	329	473	17.250	4.250	80.2%
	2	312	452	343	487	17.750	4.125	81.1%
	3	308	441	342	481	17.000	4.625	78.6%
Qwen	1	571	673	598	715	13.688	4.313	76.0%
	2	576	682	604	722	14.000	4.250	76.7%
	3	574	676	604	720	13.625	4.625	74.7%

The four controlled contrasts give the same ordering without relying on the symmetric allocation. For Llama, the gate contrasts $B - A$ and $D - C$ span 16.500–18.000 points, whereas the content contrasts $C - A$ and $D - B$ span 3.500–5.000. For Qwen, the corresponding ranges are 12.750–14.750 and 3.375–5.500 points. The interaction $D - C - B + A$ is positive but smaller: 0.500–1.500 points for Llama and 1.500–1.875 for Qwen. The gate effect is therefore larger under either content distribution in every seed.

Table 7: **Mean generation length follows the source of the stopping gate.** Values are mean generated tokens over the same 800 responses per cell.

Model	Seed	A	B	C	D
Llama	1	146.2	188.9	149.5	191.4
	2	148.4	190.1	151.0	192.2
	3	147.1	189.4	150.2	191.6
Qwen	1	153.0	193.6	154.8	194.7
	2	152.2	192.9	154.1	193.8
	3	154.1	194.0	155.6	195.1

At fixed content, replacing the standard gate with the EOS-ablated gate increases mean length by 39.5–42.7 tokens. At fixed gate, replacing the content distribution changes mean length by only 0.9–3.3 tokens. The gate also has the larger effect on endpoint accuracy (Table 6). This ordering is specific to the tested algebra task, training dose, and decoding rule. The MBPP comparison below asks whether it also holds for prefix value under a fixed continuation model.

Table 8: **The Llama d100 algebra endpoint gain is similar across the three training seeds.** Correct counts use 800 responses per seed at a 768-token cap. The shared initial checkpoint scores 599/800. Improvement compares EOS-ablated with standard SFT within each seed.

Seed	Standard SFT	EOS-ablated SFT	Added correct	Improvement (pp)
1	513	568	55	+6.875
2	527	584	57	+7.125
3	503	562	59	+7.375
Mean change			57	+7.125

Averaging the seed-level changes relative to the shared initial score gives -10.5833 points for standard SFT and -3.4583 for EOS-ablated SFT. Their difference is $+7.125$ points, used in Table 1. All three EOS-ablated checkpoints remain below the initial checkpoint, by 31, 15, and 37 correct responses.

MBPP effects across training seeds. Table 9 reports correct-count differences for Llama d100. Native endpoint differences compare each adapted checkpoint with the initial model at 768 tokens.

Source differences instead keep the initial continuation model fixed and compare the adapted-source prefixes with initial-source prefixes under the 192+576-token procedure. Each comparison uses its own initial-model reference; the table reports differences rather than absolute correct counts.

Table 9: **EOS removal improves both MBPP endpoints and prefix value in each Llama d100 seed.** Entries are differences in correct-response counts, with 4,000 responses per seed and condition. Each improvement column subtracts standard SFT from EOS-ablated SFT within the same comparison. Source contrasts use the unchanged initial continuation model throughout.

	Native endpoint difference			Fixed-continuation source difference		
Seed	Standard	EOS-ablated	Improvement	Standard	EOS-ablated	Improvement
1	-230	-86	$+144$	-253	-111	$+142$
2	-260	-112	$+148$	-246	-100	$+146$
3	-247	-91	$+156$	-243	-93	$+150$

Multiplying each correct-count improvement by $100/4000$ gives native endpoint gains of 3.60, 3.70, and 3.90 percentage points and fixed-continuation source gains of 3.55, 3.65, and 3.75 points. Their means are 3.7333 and 3.650 points. The native endpoint differences total -737 and -289 responses across seeds; multiplying these totals by $100/12000$ gives mean changes of -6.1417 and -2.4083 points. The source differences total -742 and -304, giving -6.1833 and -2.5333 points. EOS-ablated training still has negative endpoint and source contrasts relative to the respective initial-model references in every seed.

Qwen fixed-continuation effects across training seeds. Table 10 reports the corresponding source contrasts for Qwen d200. Each contrast is the adapted-source correct count minus the initial-source correct count under the same initial Qwen continuation model.

Table 10: **EOS removal improves Qwen MBPP prefix value in all three d200 seeds.** Differences use 4,000 responses per seed and condition under the fixed initial Qwen continuation model. Added correct subtracts the standard-SFT source difference from the EOS-ablated source difference.

	Source difference		Added	Gain
Seed	Standard	EOS-ablated	correct	(pp)
1	-398	-152	$+246$	$+6.150$
2	-415	-161	$+254$	$+6.350$
3	-412	-154	$+258$	$+6.450$
Mean	-408.3	-155.7	$+252.7$	$+6.317$

The standard and EOS-ablated source differences sum to -1225 and -467 responses, corresponding to three-seed means of -10.2083 and -3.8917 percentage points. EOS ablation adds 246, 254, and 258 correct responses within seed; their total of 758 gives a mean improvement of 6.3167 points. Every EOS-ablated source remains below the initial source under the fixed initial continuation model.

MBPP stopping-gate–content cross. We apply Equation 7 during the 192-token prefix stage, then use the unchanged initial checkpoint for the 576-token continuation in all four cells. Table 11 reports each cell as its correct-count difference from the initial-source condition under that same continuation model. Hence A and D reproduce the standard and EOS-ablated source differences above, while B and C exchange one prefix-generation factor. Subtracting the common initial-source reference does not change any of the four contrasts in Equation 8.

For Llama, the two gate contrasts span 2.700–2.900 points, while the two content contrasts span 0.775–0.875. For Qwen, the corresponding ranges are 4.700–5.250 and 1.050–1.450 points. The gate contrast is larger under either content distribution in every seed, with the continuation checkpoint held fixed. This ordering has not been tested under other training objectives or model scales.

MBPP task selection and prompting. We use the sanitized MBPP collection distributed through the ModelScope `opencompass/mbpp` mirror. A Python random generator with seed 42 shuffles

Table 11: **The stopping gate also has the larger effect on MBPP prefix value.** A–D are correct-count differences from the initial-source condition under the unchanged initial continuation model, with 4,000 responses per cell and seed. G and K are percentage-point allocations; gate share divides G by the percentage-point change represented by $D - A$.

Model/steps	Seed	A	B	C	D	G (pp)	K (pp)	Gate share
Llama/d100	1	−253	−142	−219	−111	2.738	0.813	77.1%
	2	−246	−134	−212	−100	2.800	0.850	76.7%
	3	−243	−127	−208	−93	2.888	0.863	77.0%
Qwen/d200	1	−398	−210	−356	−152	4.900	1.250	79.7%
	2	−415	−218	−367	−161	5.038	1.313	79.3%
	3	−412	−212	−364	−154	5.125	1.325	79.5%

the source records; we take the first 250 and sort them by numeric task ID. All model states and training conditions use this fixed set. Each task's `prompt` field follows the instruction "Complete the following Python function. Provide only the code that completes (or defines) the function." No demonstration answer is added. The resulting user message is serialized with the model's native chat template.

Each model state produces 16 sampled responses per task, or 4,000 responses in total, at temperature 0.8 and top-$p = 1.0$, with generation seed 42. The endpoint budgets are 192 and 768 generated tokens. Prefix exchange uses 192 tokens from the source checkpoint and up to 576 additional tokens from the designated continuation model. Scoring applies to the complete prefix-plus-continuation response.

MBPP extraction and execution. The extractor removes boundary Markdown code fences and parses the remaining complete text as Python. It does not search explanatory prose for an embedded program, choose among candidate code blocks, or repair invalid code. Each candidate runs in a fresh Python 3.10 subprocess, followed by the task's supplied `test_imports` and `test_list`. The wall-clock limit is three seconds, including interpreter startup, program execution, and tests; the CPU limit is five seconds. The runner attempts a 1 GiB address-space limit, but proceeds if the platform does not support setting it. The judge accepts only when subprocess standard output, after whitespace trimming, equals its `PASS` marker. Additional program output can therefore fail this strict judge even when the assertions pass. Syntax errors, runtime exceptions, failed tests, and timeouts count as incorrect.

MBPP score aggregation. C_m^L is the fraction of all 4,000 responses accepted by this judge at budget L; absent and unparseable programs remain in the denominator. It is neither partial credit over tests nor question-level pass@16. The rate W_m^{192} is the fraction of responses that terminated by token 192 and failed the judge. An unfinished program truncated at the cap is incorrect at that budget but is not an ended error unless generation had terminated. We compute MBPP endpoint and source differences within each seed, convert them to percentage points, and average over three training seeds. The shared initial checkpoint is not treated as three independent initial-model runs.

The gate–content comparison scores the complete prefix-plus-continuation program. If an already complete prefix receives a nonempty continuation, appended text can change its parseability or test outcome. The aggregate cell scores do not separate this effect from improvements on unfinished prefixes. We therefore interpret the MBPP allocation as a property of the fixed handoff and scoring procedure, not as an isolated measure of program quality before handoff.

A.2 EOS ablation on MATH-500

We evaluate the complete 500-problem test split of `HuggingFaceH4/MATH-500` (Lightman et al., 2024), with eight sampled responses per problem and 4,000 responses per model condition. Generation uses a 768-token cap, temperature 0.8, top-$p = 1.0$, and sampling enabled. Each generation chunk receives a deterministic SHA256-derived seed incorporating the base seed 42, experiment name, checkpoint, token budget, and question-ID range. Table 12 reports the initial reference, standard SFT, and EOS-ablated SFT at 50 training steps (d50) for three training seeds in

each family. Each family shares one initial-model reference across the training-seed comparisons. Accuracy is the fraction of sampled responses judged correct, not pass@8 or majority-vote accuracy.

Table 12: **EOS ablation improves MATH-500 accuracy but leaves most of the initial-model deficit.** Correct-response counts are out of 4,000 for each condition. Gain compares EOS-ablated with standard SFT within the same training seed. Mean rows average the three seed-level gains. Intervals use the paired seed differences and two degrees of freedom.

Model	Seed	Initial	Standard	EOS-ablated	Gain (pp)	95% t interval
Llama	1	1724	680	808	$+3.20$	—
	2	1724	536	688	$+3.80$	—
	3	1724	648	796	$+3.70$	—
	Mean	—	—	—	$+3.57$	$[2.77, 4.37]$
Qwen	1	2528	1592	1636	$+1.10$	—
	2	2528	1796	1836	$+1.00$	—
	3	2528	1804	1868	$+1.60$	—
	Mean	—	—	—	$+1.23$	$[0.43, 2.03]$

For training seed s, let $d_s = 100(n_{s,\mathrm{abl}} - n_{s,\mathrm{std}})/4000$ be the improvement in percentage points. The reported interval is $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$, where s_d is the sample standard deviation of the three differences and $t_{0.975,2} = 4.30265$. This gives 3.5667 ± 0.7985 points for Llama and 1.2333 ± 0.7985 for Qwen. The intervals assume independent, approximately normal seed-level differences and summarize variability across three paired training seeds on the fixed test set. With only three seeds, normality cannot be assessed reliably; these intervals do not quantify uncertainty over new questions. The shared initial reference does not enter d_s. Deficit-recovery percentages below are ratios of mean point estimates, without propagated uncertainty intervals.

Llama's initial accuracy is 43.10%, compared with mean accuracies of 15.53% for standard SFT and 19.10% after EOS ablation. Qwen's corresponding values are 63.20%, 43.27%, and 44.50%. Dividing the mean improvement by the mean initial-to-standard deficit gives reductions of 12.94% and 6.19%, respectively. EOS ablation therefore recovers only a small part of the MATH-500 adaptation deficit. These endpoint comparisons do not themselves identify a stopping-gate versus conditional-content allocation.

Eight-answer majority voting. We also aggregate the same eight sampled responses per question at the 768-token budget using majority voting (Maj@8), following answer aggregation in self-consistency (Wang et al., 2023). Answers are grouped by string or numerical equivalence, the largest group supplies the prediction, and ties are broken pseudorandomly. Each question contributes one final prediction to the 500-question score; pass@1 instead averages the correctness of all 4,000 responses. Voting uses eight responses per prediction and therefore costs more than generating a single response. Its paired question-bootstrap intervals concern the Maj@8 difference; response accuracy uses the separate seed-level intervals above. Initial-model outputs are shared across adapted seeds and reported once per family.

Table 14 compares voted answers on the same questions. An EOS win means that only the EOS-ablated prediction is correct; a standard win means that only the standard-SFT prediction is correct. A tie means both are correct or both are incorrect, irrespective of whether the predicted strings match. The difference is $100(n_{\mathrm{EOS\ win}} - n_{\mathrm{standard\ win}})/500$ percentage points. The reported 95% bootstrap intervals resample paired questions within each seed; they differ from the three-seed t intervals for response accuracy above.

Averaged across training seeds, Maj@8 rises from 16.73% to 22.87% for Llama and from 53.60% to 55.87% for Qwen, giving gains of 6.13 and 2.27 points. The Llama intervals and the Qwen seed-1 interval exclude zero; the Qwen seed-2 interval reaches zero and seed 3 crosses it. Although EOS ablation improves Maj@8 in every seed, all adapted checkpoints remain below their initial models. The additional benefit of voting itself need not increase: Qwen seed 3 gains 9.90 points from voting after EOS ablation, compared with 10.30 under standard SFT. These results do not by themselves identify changes in response diversity or the stopping gate.

Table 13: **EOS-ablation gains persist under eight-answer voting on MATH-500.** All adapted checkpoints use d50. Pass@1 reuses the response counts in Table 12; Maj@8 counts correct voted answers out of 500. Voting gain subtracts pass@1 from Maj@8 within each row, in percentage points.

Model	Condition	Seed	Pass@1 (%)	Maj@8 count	Maj@8 (%)	Voting gain (pp)
Llama	Initial	—	43.10	250	50.0	+6.90
	Standard	1	17.00	93	18.6	+1.60
	Standard	2	13.40	71	14.2	+0.80
	Standard	3	16.20	87	17.4	+1.20
	EOS-ablated	1	20.20	121	24.2	+4.00
	EOS-ablated	2	17.20	99	19.8	+2.60
	EOS-ablated	3	19.90	123	24.6	+4.70
Qwen	Initial	—	63.20	333	66.6	+3.40
	Standard	1	39.80	251	50.2	+10.40
	Standard	2	44.90	276	55.2	+10.30
	Standard	3	45.10	277	55.4	+10.30
	EOS-ablated	1	40.90	267	53.4	+12.50
	EOS-ablated	2	45.90	288	57.6	+11.70
	EOS-ablated	3	46.70	283	56.6	+9.90

Table 14: **Every seed has a positive voting point estimate; the interval for Qwen seed 2 reaches zero and that for seed 3 crosses it.** Paired correctness outcomes for Maj@8 on 500 questions. Intervals are the reported question-level paired bootstrap intervals, in percentage points; no across-seed interval is inferred from them.

Model	Seed	EOS wins	Ties	Standard wins	Gain (pp)	95% interval (pp)
Llama	1	39	450	11	+5.6	[2.2, 9.0]
	2	35	458	7	+5.6	[2.2, 9.0]
	3	46	444	10	+7.2	[3.6, 10.8]
Qwen	1	27	462	11	+3.2	[0.6, 5.8]
	2	24	464	12	+2.4	[0.0, 4.8]
	3	20	466	14	+1.2	[−1.0, 3.4]

A.3 EOS ABLATION ON HUMANEVAL

We evaluate Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct on all 164 HumanEval tasks (Chen et al., 2021), using the OpenCompass distribution of the `openai_humaneval` test parquet without filtering tasks. For each family and training seed, the standard and EOS-ablated checkpoints are the same d50 checkpoints used for MATH-500: 50 SFT steps on the 500-example GSM8K training subset. No additional training or code-domain fine-tuning is performed. Each condition contains eight randomly sampled responses per task, or 1,312 responses, with a 768-token generation cap, temperature 0.3, top-$p = 0.95$, and sampling enabled. These settings apply to all three conditions and differ from the temperature-0.8 prefix-exchange protocol. Standard and EOS-ablated checkpoints are compared within each of three training seeds; the initial-model reference is shared across seeds within each family.

Generated code is processed by an AST-based code extractor and evaluated against the HumanEval functional tests in an execution sandbox with a four-second test-execution limit. Extraction failures, syntax errors, failed assertions, and timeouts all count as incorrect responses. Accuracy is the fraction of responses passing all tests. With eight responses per task, this is the mean estimated pass@1 across tasks, not pass@8. Table 15 reports the endpoint counts; Table 16 reports syntax errors and timeouts separately.

For each seed, $d_s = 100(n_{s,\mathrm{abl}} - n_{s,\mathrm{std}})/1312$. As in Appendix A.2, the interval is $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$. Llama's mean, sample standard deviation, and standard error are 6.6819, 0.1918, and 0.1107 points; Qwen's are 6.0722, 0.1587, and 0.0916. All calculations use the integer counts before rounding. These intervals have the same fixed-benchmark scope and three-seed distributional assumptions as the MATH-500 intervals; they do not quantify uncertainty over new tasks. The initial reference does not enter these differences, so the intervals are not tests of improvement over the initial model.

Table 15: **EOS ablation improves HumanEval functional correctness and exceeds the initial reference in every tested seed.** Counts are out of 1,312 sampled responses per condition at 768 tokens. Gain compares EOS-ablated with standard SFT. The initial reference is shared within each family. Mean rows report paired seed-level t intervals with two degrees of freedom.

Model	Seed	Initial	Standard	EOS-ablated	Gain (pp)	95% t interval
Llama	1	802	748	836	+6.71	—
	2	802	741	831	+6.86	—
	3	802	755	840	+6.48	—
	Mean	—	—	—	+6.68	[6.21, 7.16]
Qwen	1	985	932	1011	+6.02	—
	2	985	940	1022	+6.25	—
	3	985	926	1004	+5.95	—
	Mean	—	—	—	+6.07	[5.68, 6.47]

Table 16: **Fewer syntax errors accompany the HumanEval gains.** Each condition contains 1,312 responses. Initial counts are shared across seeds. All conditions have zero observed timeouts. Other functional-test failures remain incorrect in Table 15.

		Syntax errors			Timeouts		
Model	Seed	Initial	Standard	EOS-ablated	Initial	Standard	EOS-ablated
Llama	1	68	135	51	0	0	0
	2	68	142	56	0	0	0
	3	68	128	48	0	0	0
Qwen	1	41	82	25	0	0	0
	2	41	76	21	0	0	0
	3	41	88	28	0	0	0

EOS ablation adds 85–90 correct responses for Llama and 78–82 for Qwen relative to standard SFT. It also exceeds the shared initial reference by 29–38 and 19–37 responses, respectively. Syntax-error counts decrease in every seed, and no timeout is observed under the stated limit. The gains concern functional correctness under the tested execution conditions; they neither guarantee termination on other inputs nor separate stopping and content effects. The analysis uses aggregate counts, without a question-level bootstrap.

A.4 Local handoff diagnostic on MATH-500

We compare both prefix-exchange directions between the initial Instruct checkpoint and its SFT adaptation in each family on MATH-500. Initial refers to Llama-3.1-8B-Instruct or Qwen2.5-7B-Instruct, not the corresponding pretrained base model. This diagnostic uses greedy decoding with runtime seed 42, distinct from the sampled endpoint and prefix-exchange evaluations. The receiver scores both the exchanged and native trajectories throughout, including the prefix, using token-level negative log-likelihood (NLL, natural logarithms). Each token is conditioned on the question and preceding tokens. The receiver is unchanged across the two conditions in each comparison.

For a paired question i, let

$$J_{c,i} = \frac{1}{16}\sum_{t=193}^{208} \ell_{c,i,t} - \frac{1}{16}\sum_{t=177}^{192} \ell_{c,i,t}, \quad c \in \{\text{swap}, \text{native}\},$$
$$\delta_i = J_{\text{swap},i} - J_{\text{native},i}, \qquad \rho_{\text{splice}} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \delta_i\right). \tag{10}$$

Here $\ell_{c,i,t}$ is the receiver's NLL for token t. The ratio ρ_{splice} compares the post/pre perplexity ratio in exchanged trajectories with that in native trajectories. The native change provides a reference for variation with token position; it does not isolate a pure interface effect, since prefix content and subsequent generation paths also differ.

Each comparison starts from 500 questions and includes a pair only when both trajectories cover the complete windows through token 208, without filtering by answer correctness. In the order shown in Figure 3, the retained counts are 472, 456, 482, and 465. Exclusions due to termination at or before token 192 number 19, 31, 12, and 24; a further 9, 13, 6, and 11 pairs are excluded for termination at tokens 193–207. These categories account for all 500 questions in each comparison. Because the retained populations differ, row-to-row differences do not isolate a receiver effect. The diagnostic applies only to the retained pairs, excluding early-ending responses.

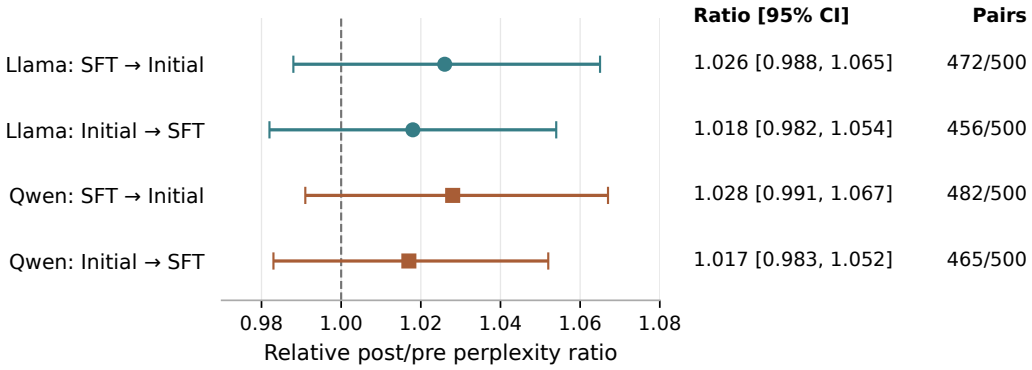

Figure 3: **Mean excess handoff changes are small on window-complete MATH-500 pairs.** Points show ρ_{splice} from Equation 10; bars are the reported marginal 95% bootstrap intervals from 10,000 question-level paired bootstrap draws. Every row compares an exchange with its receiver-to-receiver native control. The dashed line marks a ratio of one, and the right column gives retained pairs out of 500 questions. These are window-averaged diagnostics, not tokenwise curves.

Mean excess changes are 0.026 and 0.018 nats/token for Llama and 0.028 and 0.017 for Qwen, ordered by initial and adapted receiver. Their relative perplexity ratios are 1.026, 1.018, 1.028, and 1.017; the corresponding interval upper bounds are 1.065, 1.054, 1.067, and 1.052. The reported medians of δ_i range from 0.013 to 0.022 nats/token and the 95th percentiles from 0.124 to 0.155. These window averages are small, but can hide individual token spikes. They neither establish tokenwise smoothness nor rule out other distribution shifts or effects under sampled decoding. The distinct gate–content policy exchange is outside the scope of this handoff diagnostic.

B Prefix-Exchange Results

Protocol and population. The exchange experiment retains all 800 outputs per prefix source, including terminated prefixes. A stored prompt and prefix token sequence is replayed unchanged under both continuation checkpoints, using the same tokenizer within a model family and no handoff instruction. The exchange restarts continuation randomness and batch layout. At 768 tokens, its initial diagonal differs from native single-pass generation by $+0.875$ points and its adapted diagonals by $+0.250$, -1.125, and -0.125. All factorial contrasts use the two-stage cells in Table 17. Their prefix-correct counts are 181 for the initial model and 274, 286, and 269 for the adapted models, matching the independent 192-token counts.

Table 17: **Correct-response counts for algebra prefix exchange.** Counts correspond to the C_{ij} scores, with 800 outputs per cell. The first subscript is the prefix source and the second the continuation checkpoint. 0 is the initial model and t its adaptation. The initial-to-initial cell is shared, not rerun for each seed.

Training seed	C_{00}	C_{0t}	C_{t0}	C_{tt}	$(C_{tt} - C_{00})$ (pp)
1	606	577	524	515	-11.375
2	606	581	528	518	-11.000
3	606	581	499	502	-13.000

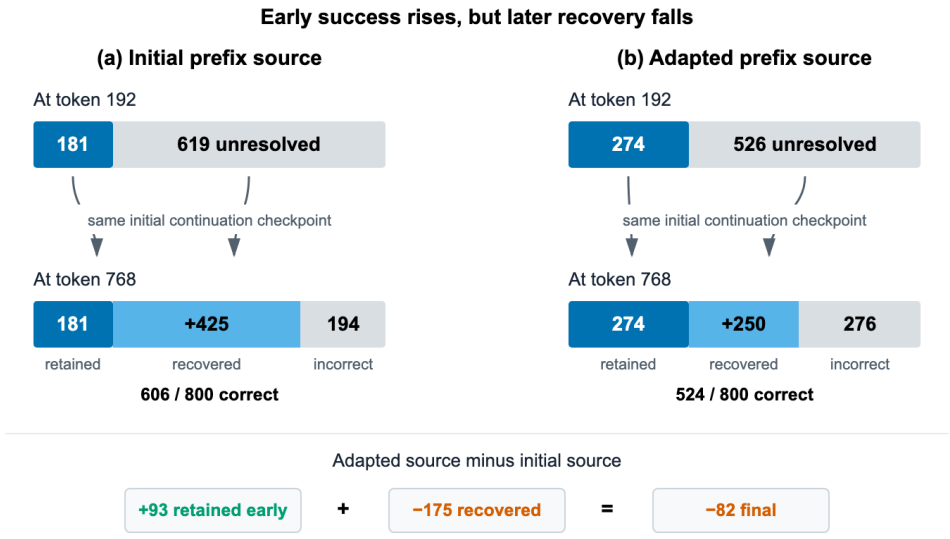

Figure 4: **The seed-1 reversal is a loss of later recoveries, not a loss of early successes.** Both prefix sources are continued by the same initial checkpoint. Adapted prefixes contain 93 more correct answers at token 192, all of which are retained, but they yield 175 fewer recoveries and 82 fewer final successes. Table 18 gives the same accounting for both continuation checkpoints and all tested seeds.

Table 18: **On algebra, early correctness and later recovery move in opposite directions.** Each entry is the adapted-source count minus the initial-source count, with the continuation checkpoint fixed. ΔP is the prefix-correct difference. ΔU and ΔR are the retained-success and recovered-success terms in Equation 5; their sum is the final prefix-source effect ΔC. Every comparison contains 800 responses per source; the GSM8K rows concern the original 50-question diagnostic subset.

			Initial continuation			Adapted continuation		
Task	Model / seed	ΔP	ΔU	ΔR	ΔC	ΔU	ΔR	ΔC
Algebra	Llama / 1	$+93$	$+93$	-175	-82	$+93$	-155	-62
	Llama / 2	$+105$	$+105$	-183	-78	$+105$	-168	-63
	Llama / 3	$+88$	$+88$	-195	-107	$+88$	-167	-79
GSM8K	Llama / 1	-43	-42	-16	-58	-39	-17	-56
	Qwen / 1	-4	-7	$+14$	$+7$	-4	$+5$	$+1$

Replacement order. The total difference can be written in either order:

$$C_{tt} - C_{00} = (C_{t0} - C_{00}) + (C_{tt} - C_{t0}), \tag{11}$$
$$= (C_{0t} - C_{00}) + (C_{tt} - C_{0t}). \tag{12}$$

The interaction $J = C_{tt} - C_{t0} - C_{0t} + C_{00}$ is $+2.500$, $+1.875$, and $+3.500$ points. The contribution assigned to the prefix depends on which factor is replaced first, so we report both controlled contrasts. Averaging the two orders gives prefix differences of -9.000, -8.8125, and -11.625 points and continuation differences of -2.375, -2.1875, and -1.375. This averaging is a descriptive allocation, not a unique causal mediation fraction.

Question-level variation. The 50-row table contains ten strict correct counts per question, with 16 responses in each cell. Summing it reproduces all counts in Table 17. A post-hoc question bootstrap carries the full ten-column row together and averages the three adapted seeds within each draw, with the shared initial cell included once. Using the resampling procedure in Appendix A, the mean prefix contrast is -11.125 points with the initial continuation model (95% interval $[-15.79, -6.21]$) and -8.500 with the adapted continuation model ($[-12.88, -3.75]$). The corresponding continuation con-

trasts are -3.292 on initial prefixes ($[-4.96, -1.79]$) and -0.667 on adapted prefixes ($[-1.71, 0.37]$). These intervals condition on the tested models, seeds, and questions.

To compare the two stages, each bootstrap draw computes the smaller absolute prefix contrast minus the larger absolute continuation contrast. The observed margin is 5.208 points; it is positive in 98.35% of the 10,000 question-bootstrap draws, with a percentile interval of $[0.42, 9.71]$. Applying the analogous definition to the empirically larger stage on the original 50-question GSM8K diagnostic subset gives a Llama prefix margin of 6.750 points (98.43% positive; $[0.50, 11.50]$) and a Qwen seed-1 continuation margin of 3.375 points (85.98% positive; $[-1.50, 7.25]$). Qwen seeds 2 and 3 have aggregate continuation margins of 3.750 and 3.875 points, respectively; subset bootstrap intervals are not reported for these two seeds. These subset intervals do not apply to the full-test counts in Table 19.

Interpretation of unfinished responses. The full comparison includes both terminated and still-active prefixes. The reported answer transitions show no loss of correct prefix answers in any of the ten continuation conditions. Recovery rates restricted to initially incorrect or still-active outputs compare source-dependent subsets: adaptation can change which responses enter them. Such rates do not isolate damage to a fixed set of unfinished reasoning states or establish a shared cause for the early advantage and later loss. The full-population exchange instead identifies the larger effect of changing the prefix source under the handoff procedure. Among seed-1 unterminated initial-model prefixes, 427 final successes comprise two prefixes already correct at token 192 and 425 initially incorrect prefixes that recover; the two counts refer to different transition subsets.

Extraction sensitivity. The factorial contrasts can be recomputed from the strict question-level counts. Reported expanded-minus-strict changes reach approximately 0.6 points. These summaries do not determine the exact expanded-score contrasts or exclude a role for formatting.

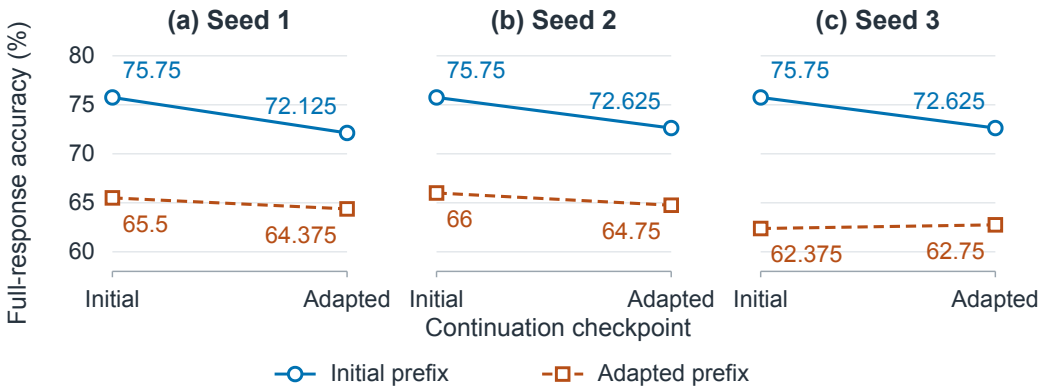

Figure 5: **Changing the prefix source matters more than changing the model that continues it.** Each panel crosses the initial and adapted Llama checkpoints for one training seed. The upper line starts from initial-model prefixes; the lower line starts from adapted prefixes. Every point uses 800 full responses under the same two-stage generation procedure. The initial-to-initial point is shared across panels.

C Held-Out GSM8K Prefix Exchange

Questions and generation. The primary comparison evaluates all 1,319 questions in the GSM8K test split with four responses per question, giving 5,276 outputs per cell. Evaluation uses Llama d100 seed 1 and all three Qwen d100 adapters trained on the same 500 training-split examples, with no additional fine-tuning. Each prefix is reused under both continuation checkpoints. The initial-to-initial collection is shared across Qwen seeds, giving four distinct Llama cells and ten Qwen cells, or 73,864 full-response records in total.

Each original question is rendered using the model's chat template with a generation prompt, without additional format instructions. Sampling retains seed 42, temperature 0.8, and top-$p = 1$. A prefix

contains up to 192 generated tokens. If the source terminates earlier, its generated tokens before the terminator are retained and the record remains in the population. Continuation receives the original prompt IDs and retained prefix IDs directly, with no decode-and-retokenize step or inserted instruction, and has a 576-token output cap. Previously terminated outputs can yield empty or nonempty continuation. Text for scoring concatenates the decoded prefix and continuation; an appended answer marker or answer cue can change which number is extracted. The full-test contrasts therefore include answer replacement under this handoff procedure. The cell totals alone do not separate its contribution among previously terminated prefixes.

Numerical extraction. The primary C_{num} rule first extracts the first number after the last #### marker; otherwise it uses a number near the last answer cue, within a 40-character window and starting at offset at most 12; otherwise it takes the final number in the response. Gold answers come from the dataset's answer field. Commas and currency signs are normalized, and comparison uses tolerance $10^{-6}\max(1, |\text{gold}|)$. C_{marker} retains the marker and answer-cue rules but omits the final-number fallback; C_{last} uses only the final number. The three rules score the same complete responses in the original 50-question extraction-sensitivity analysis; the full-test comparison reports the primary numerical score.

Table 19: **Correct-response counts for full-test GSM8K prefix exchange.** Each cell contains 5,276 outputs from 1,319 questions with four samples each. C_{00} is shared across training seeds within a model family. Llama has one adaptation seed; Qwen has three.

Model	Seed	C_{00}	C_{0t}	C_{t0}	C_{tt}
Llama	1	3983	3972	3607	3618
Qwen	1	4386	4642	4435	4663
	2	4386	4628	4417	4651
	3	4386	4655	4428	4670

On the full test set, changing the prefix source with the initial/adapted continuation model fixed gives $-7.127/-6.710$ points for Llama. For Qwen, the corresponding effects are $+0.929/+0.398$, $+0.588/+0.436$, and $+0.796/+0.284$ points across seeds. Changing the continuation model on initial/adapted prefixes gives $-0.208/+0.208$ for Llama and $+4.852/+4.321$, $+4.587/+4.435$, and $+5.099/+4.587$ for Qwen. The interactions are $+0.417$ points for Llama and -0.531, -0.152, and -0.512 for Qwen. Table 20 reports question-level paired percentile bootstrap intervals for the full-test stage-ordering margins; the four individual contrasts are reported as point estimates. The intervals for the original 50-question diagnostics below concern that subset alone.

Full-test uncertainty. We use 10,000 paired question-bootstrap draws with resampling seed 20260828. Each draw samples 1,319 question IDs with replacement and retains all four responses and compared cells for each sampled question. Cell accuracies are pooled response means; the stage-ordering margin is recomputed within each draw. The 95% interval takes the 2.5th and 97.5th percentiles of these bootstrap margins, without bias correction or acceleration.

Table 20: **The full-test stage ordering is supported in each comparison.** Margins and intervals are in percentage points, using all 1,319 GSM8K test questions. $D_P = \min(|P_0|, |P_t|) - \max(|K_0|, |K_t|)$ and $D_K = \min(|K_0|, |K_t|) - \max(|P_0|, |P_t|)$. Here $P_0 = C_{t0} - C_{00}$, $P_t = C_{tt} - C_{0t}$, $K_0 = C_{0t} - C_{00}$, and $K_t = C_{tt} - C_{t0}$. Question-level paired bootstrap retains each question's four responses and all compared cells together. Intervals are marginal 95% percentile intervals for the individual comparisons.

Model / seed	Ordering margin	Estimate (pp)	95% interval
Llama / 1	D_P	6.50	$[5.36, 7.34]$
Qwen / 1	D_K	3.39	$[2.41, 4.08]$
Qwen / 2	D_K	3.85	$[2.91, 4.53]$
Qwen / 3	D_K	3.79	$[2.82, 4.47]$

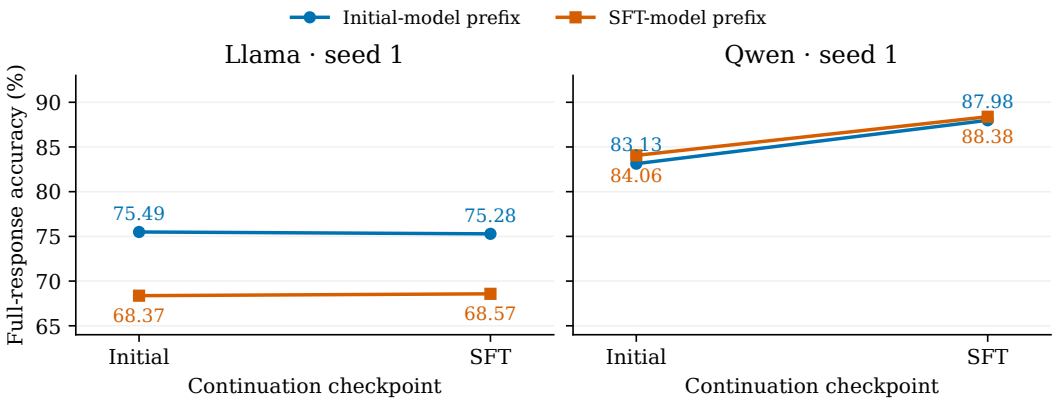

Figure 6: **Adaptation affects different parts of generation on GSM8K in seed 1.** For Llama, changing the prefix source separates the two curves, while changing the continuation model has little aggregate effect. For Qwen, both prefix sources benefit from the adapted continuation model. Each point uses 5,276 outputs on all 1,319 test questions, with four samples per question. Table 19 reports the two additional Qwen seeds.

C.1 Diagnostics on a 50-question subset

The remaining GSM8K analyses use a 50-question diagnostic subset with 16 responses per question to examine answer extraction, retention, and recovery in detail. First-error correction and minimum-length decoding use the same subset (Appendices H and F). These scoring and intervention analyses are separate from the full-test stage comparison. The subset was selected by shuffling test indices with Python's `random.Random(20260828)`, taking the first 50, and sorting the selected indices. No exact question-text matches with the training split were reported; this does not establish absence of benchmark exposure during pretraining. The seed-1 prefix-correct counts are 519/476 for initial/adapted Llama and 194/190 for Qwen. Table 21 retains the subset's seed-1 numerical scores as the reference for its extraction and trajectory diagnostics.

Single-pass reference. For seed 1, single-pass 768-token numerical correct counts are 606/525 for initial/adapted Llama and 667/702 for Qwen. The two-stage diagonal differences are therefore $-0.250/+2.875$ points for Llama and $-0.250/+0.500$ for Qwen. Native generation is a separate reference, not a replacement for either cell of the factorial comparison.

Table 21: **Answer-extraction sensitivity on the original 50-question seed-1 GSM8K subset.** Entries are correct counts out of 800. Omitting the final-number fallback reverses Llama's training comparison; final-number extraction preserves the main numerical ordering.

Model	Rule	C_{00}	C_{0t}	C_{t0}	C_{tt}	Gain (pp)
Llama	Numeric	604	604	546	548	-7.000
	Marker	59	72	250	254	$+24.375$
	Last number	604	604	546	548	-7.000
Qwen	Numeric	665	705	672	706	$+5.125$
	Marker	2	65	5	7	$+0.625$
	Last number	669	707	674	708	$+4.875$

Effect of the extraction rule. On these same subset responses, Numeric and Last number both give a Llama training difference of -7.000 points, whereas Marker gives $+24.375$. Marker accepts a number after either a `####` marker or an answer cue; it rejects responses without either, even if the final number is correct. Because the prompt does not require these cues, its reversal reflects sensitivity to answer presentation as well as correctness. The negative numerical-score difference persists under both rules that allow unmarked final answers. This check assesses extraction sensitivity on the subset; it is not a manual validation of every extracted answer.

Retention and recovery. Continuation can lose prefix successes on GSM8K, unlike the inspected algebra exchange. For seed 1, in cell order 00, $0t$, $t0$, tt, Llama retains 512/513/470/474 successes, loses 7/6/6/2, and recovers 92/91/76/74 initially incorrect outputs. Qwen retains 190/190/183/186, loses 4/4/7/4, and recovers 475/515/489/520. Retained plus recovered counts reproduce every final numerical score. On initial Qwen prefixes, the adapted continuation model gains 40 outputs entirely through additional recoveries. On adapted prefixes, its 34-output gain comprises 31 additional recoveries and three fewer lost successes.

Stopping strata and interpretation. For seed 1, initial/adapted Llama terminate within the prefix allowance in 584/699 records; Qwen does so in 147/97. Accuracy conditional on these source-dependent groups does not compare a common unfinished population. Holding the still-active initial Qwen prefixes fixed, switching continuation models increases the correct count from 535 to 569 out of 653. On the still-active adapted prefixes, it increases from 587 to 620 out of 703. Llama's corresponding changes are 134 to 132 out of 216 and 65 to 69 out of 101. Qwen also benefits from the adapted continuation model on each fixed set of unterminated prefixes. The small Llama differences do not establish equal continuation performance.

The pre-experiment protocol recorded both trajectory-shape and stage-ordering predictions for the original seed-1 GSM8K comparison. The predicted short-budget Llama gain and shrinking Qwen loss did not occur, whereas the predicted larger Llama prefix effect and larger Qwen continuation effect did. The extended Qwen comparison shows the same stage ordering in all three training seeds. Task content, answer form, and prompt constraints change together between algebra and GSM8K. The results therefore do not identify format instructions or training-task matching as necessary conditions for reversal. They extend the comparison to one held-out mathematical task, with one Llama seed and three Qwen seeds.

C.2 Question-level scope of the stage comparison

The following deletion analysis concerns the original 50-question algebra and GSM8K collections, not the full-test GSM8K comparison. For the five comparisons with question-level tables, we remove each question in turn, retain all 16 samples for every remaining question and cell, and recompute the four contrasts. The same stage has the larger effect after each of the 50 deletions in every comparison. Table 22 reports the smallest remaining margin between the smaller absolute contrast for the dominant stage and the larger absolute contrast for the other stage. The 1.786-point minimum belongs to Qwen seed 1 on the GSM8K subset. A positive deletion margin establishes stability to removing one observed question, not statistical significance over new questions. This deletion analysis does not include Qwen seeds 2 and 3.

Table 22: **No single question determines the stage ordering in the original 50-question collections.** Minimum leave-one-question-out margins are in percentage points over 784 responses per cell. The last two columns count questions where each stage's smaller absolute contrast strictly exceeds the other stage's larger absolute contrast. All remaining questions have ties or overlapping contrast ranges.

Task	Model / seed	Minimum margin	Prefix larger	Continuation larger
Algebra	Llama / 1	3.444	32/50	2/50
	Llama / 2	3.954	25/50	1/50
	Llama / 3	6.122	29/50	0/50
GSM8K	Llama / 1	5.740	33/50	0/50
	Qwen / 1	1.786	10/50	3/50

Qwen seed 1 has a larger aggregate continuation effect, yet only three questions individually meet the strict continuation-dominance criterion. Aggregate effect size therefore does not describe how often one stage dominates across questions. The reported ordering concerns the model–task average.

D CONTINUATION INSTRUCTIONS AND ANSWER SCORING

Changing the continuation instruction. At the nominal 192-token cutoff, three alternatives use an additional allowance of 64 output tokens: A, appending `Final answer now:`; B, appending `Continue solving, then give the final answer:`; and B', bare continuation. Sampling uses temperature 0.8 and top-$p = 1$. Under the reported protocol, bare continuation replays a segment of the native trajectory. Some responses terminate before the cutoff; the early-termination analysis below explains how an appended instruction can reopen them.

Each branch is scored on the original prefix plus its continuation. The strict extractor requires the prescribed answer form; the expanded extractor accepts additional algebra-answer forms. Both rules are fixed across branches. The A versus B' comparison holds the additional output cap fixed, but not input length, actual computation, or latency. Either branch can compute further before producing an answer. The 64-token extension may constrain further reasoning differently across instructions, so these comparisons describe the instruction and cap jointly, rather than an instruction effect independent of the output allowance.

D.1 ANSWER ELICITATION AND TRAINING STATE

We test whether prompting a model to finish its answer changes the relative performance of its initial and adapted checkpoints. Define

$$E_t^e = C_t^{A,e} - C_t^{B',e}, \qquad I^e = \frac{1}{3}\sum_{s=1}^{3} E_{100,s}^e - E_0^e, \tag{13}$$

where s indexes the three adapted training seeds. E is the effect of answer-now relative to bare continuation; I is its change after adaptation, averaged over the observed seeds. Neither quantity measures whether the model already represented the answer internally.

Qwen's adapted checkpoints benefit from answer-now. Under strict extraction, $A - B'$ is $+3.375$ points for the initial Qwen model and $+6.750$, $+8.375$, and $+9.000$ for the adapted checkpoints. Expanded extraction changes the initial-model difference to -0.125 points, while the three adapted differences remain $+5.625$, $+7.750$, and $+7.625$. The result therefore cannot be explained solely by requiring a literal `ANSWER:` marker in the strict evaluator.

The mean interaction is $I^{\text{strict}} = 4.67$ points, with a paired question-bootstrap interval of $[2.04, 7.38]$, and 7.13 under expanded extraction, with interval $[3.92, 10.54]$. Resampling retains each question's outputs and all three adapted states together, including the shared initial checkpoint once. The intervals concern the observed checkpoints, not variation over model families or new training seeds. Restricting the analysis to the 605 question–sample identifiers that remain unterminated before token 192 in every Qwen state leaves the interaction positive: 5.79 points under strict extraction and 9.20 under expanded extraction. Thus differences in early-termination composition do not alone explain its direction in this collection.

Llama does not show the same interaction. The initial Llama model reaches 44.0% under answer-now and 47.5% under bare continuation. Its three adapted checkpoints also favor bare continuation, by 3.75–4.63 points under strict extraction. The mean interaction is -0.75 points, with interval $[-3.29, 1.83]$; expanded extraction gives -0.46 points, with interval $[-3.00, 2.00]$. Thus the intervention does not clearly distinguish the two Llama training states, despite their strong budget reversal. A common answer-elicitation explanation for both models is not supported by this comparison.

Why full-response scoring changes the interpretation. Scoring only the newly generated segment would give the initial Llama model 24.625% under answer-now, just two points above its 192-token score. That segment-only increase combines 167 responses that were incorrect in the prefix but correct in the new segment with 151 responses that were correct in the prefix but not repeated correctly in the new segment. It compares different scoring windows. Complete-response scoring gives 44.0%, or a 21.375-point increase over the prefix score. The comparison with bare continuation remains negative.

Continuation instructions also affect answer retention. For initial Qwen, the instructed-continue branch B loses 78 strict prefix successes; expanded extraction reduces that count to seven. The three adapted checkpoints lose 44/46/34 under strict extraction and 6/10/6 under expanded extraction. Bare continuation loses no Qwen prefix successes. Adding a continuation instruction therefore changes more than the available output budget. Table 23 reports full-response scores; further retention counts appear below.

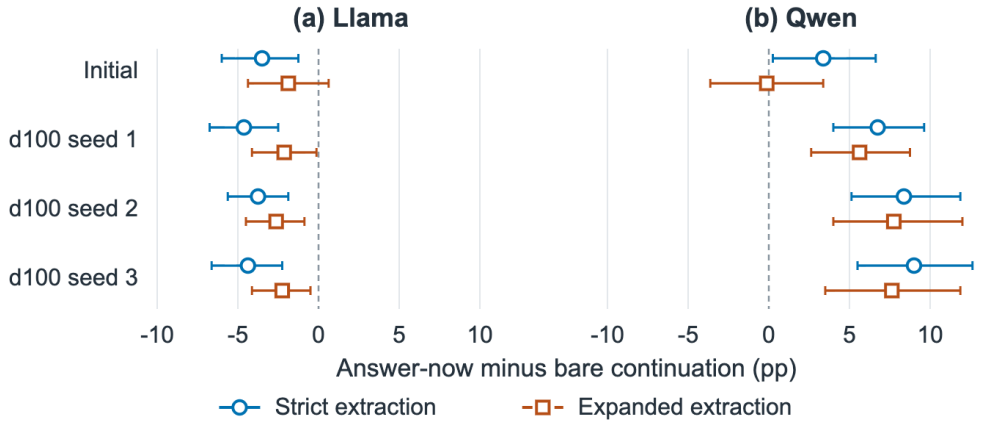

Figure 7: **An answer-now instruction helps adapted Qwen relative to bare continuation.** Points show E_t, using complete-response scores and a common 64-token extension allowance. Intervals are 95% paired question-bootstrap intervals, conditional on the displayed checkpoint and training seed. Expanded extraction changes the initial Qwen estimate from positive to near zero; adapted checkpoints retain positive estimates. Llama's point estimates favor bare continuation.

Table 23: **Complete-response branch scores.** Percent correct, each from 800 outputs. A requests an immediate answer, B requests continued solving, and B' continues without an instruction. Every score includes the original prefix.

		Strict			Expanded		
Model	State	A	B	B'	A	B	B'
Llama	Initial	44.000	45.375	47.500	49.375	49.625	51.250
	Seed 1	48.000	47.125	52.625	52.375	49.375	54.500
	Seed 2	50.500	48.875	54.250	53.375	51.125	56.000
	Seed 3	47.625	47.500	52.000	51.250	49.000	53.500
Qwen	Initial	50.125	39.875	46.750	53.125	51.625	53.250
	Seed 1	36.625	26.625	29.875	40.125	35.625	34.500
	Seed 2	37.125	26.375	28.750	41.375	34.500	33.625
	Seed 3	37.750	27.125	28.750	41.250	34.375	33.625

Where long-budget errors first appear. We truncate each native 768-token response at 192 tokens and classify that prefix as correct, containing a complete incorrect pair, or lacking a complete pair. Completeness means that both nonempty x and y components occur in the stored parsed value; we retain the recorded correctness labels. Under strict extraction, Table 24 shows that all additional native-stream successes come from the last category. Among responses lacking a complete pair at 192, the final-error counts are 166 for the initial model and 222, 221, and 225 for the adapted checkpoints. The mean increase in this group is 7.08 points, with a paired question-bootstrap interval of $[3.79, 10.33]$. The increase among complete incorrect pairs is 3.50 points, with interval $[-0.25, 6.63]$. Together they give the 10.58-point overall error increase, with interval $[5.67, 15.21]$.

Expanded extraction can read pairs from intermediate calculations: 80 of the initial model's 154 early incorrect pairs subsequently become correct. Under this extractor, the increase in final errors without a complete pair at 192 is 4.75 points, with interval $[1.67, 7.88]$, out of a total increase of 10.38

Table 24: **Llama's long-budget loss extends beyond early complete wrong answers.** Counts from 800 native 768-token responses per checkpoint under strict extraction, with the early scores obtained by truncation at 192. "No pair" includes partial answers; "recovered" counts this category's successes at 768. Previously correct responses remain correct, and complete wrong pairs do not recover in these native streams.

State	Correct at 192	Wrong pair at 192	No pair at 192	Recovered	Correct at 768
Initial	181	35	584	418	599
Seed 1	275	65	460	238	513
Seed 2	286	52	462	241	527
Seed 3	269	72	459	234	503

points. Thus the strict no-recovery observation should not be applied to every intermediate wrong pair. Both partitions describe the composition of errors; they are not causal mediation estimates because category membership can change after training.

First verifiable content errors in native failures. This supplementary diagnosis classifies the first verifiable content error in algebra responses from the initial Llama checkpoint and d100 seed 1. These categories are separate from the task correctness scores used in prefix exchange and the EOS gate–content decomposition. A verifiable error is an incorrect interpretation or equation setup, an invalid algebraic transformation, an incorrect numerical calculation, or an incorrect answer transcription. Missing answer markers, incomplete expressions, and omitted reasoning steps alone do not constitute content errors. The first demonstrably incorrect statement or operation determines the category. Responses with no explicit error or an undetermined classification are recorded separately. An error counts as early only when its incorrect content has appeared completely before the reported cutoff. The same criterion is used across checkpoints and final outcomes.

A rule-based parser checks mathematical claims line by line against the reference solution and locates the first verifiably false claim, recording its token position and a draft category. One annotator reviewed all 488 final failures (201 initial and 287 adapted), recording reasons for 28 category changes and manually resolving six cases the parser could not locate. No LLM was used for annotation. This single-reviewer procedure provides no inter-annotator agreement estimate and does not control for reviewer-expectation bias.

Table 25 covers the 201 and 287 final failures, respectively, among 800 native responses per checkpoint at a 768-token cap. Each response contributes once, according to its first verifiable error or the absence of a located error. This classification concerns intermediate reasoning, whereas Table 24 concerns whether the parser has extracted a complete answer pair.

Table 25: **Additional native failures concentrate in algebraic transformations.** First-error categories for Llama on algebra, initial checkpoint versus d100 seed 1. The timing columns retain the recorded "before 192" classification; dashes denote inapplicable timing.

	Initial		d100 seed 1	
First error type	Total	Before 192	Total	Before 192
Problem interpretation or equation setup	0	0	3	2
Transposition, elimination, or substitution	178	126	251	219
Numerical calculation	21	13	30	25
Final-answer copying or expression	1	0	2	2
No explicit content error found	1	—	1	—
Undetermined	0	—	0	—
All final failures	201	139	287	248

The timing counts partition the increase in final errors: first errors reported before 192 increase by 109, later located errors decrease from 61 to 38, and one failure per checkpoint has no explicit content error. Thus $109 - 23 = 86$. The algebraic-transformation category contributes $73/86 = 84.9\%$ of the net count increase; this is a composition calculation, not a causal attribution. Its share of all final

failures is similar across checkpoints, 88.6% versus 87.5%. The early fraction within this category rises from 70.8% to 87.3%, conditional on final failure.

The failure categories alone cannot distinguish changes in error prevalence from changes in success after an error. Table 26 adds early-error counts for successful responses. The checkpoint-specific failure populations differ, and a token cutoff need not align the number of completed reasoning operations. These summaries also do not track whether individual errors are later repaired.

This analysis covers one algebra seed. Category totals are checked against the native final scores of 599 and 513 correct responses. Matching the totals checks for omissions and double counting; it does not validate the individual categories.

Early errors across all final outcomes. The supplementary annotations report an early content error in 39 of the initial model's 599 finally correct outputs and 18 of the adapted model's 513 finally correct outputs, using the same content-based rubric. Combining these counts with the failure-side counts yields Table 26. We report only early-error presence for these successful outputs, without an error-type breakdown.

Table 26: **Early errors are more frequent after adaptation, and fewer error-bearing outputs finish correctly.** Native algebra, Llama initial versus d100 seed 1, 800 outputs per checkpoint. "Early error" denotes a reported content error before the nominal 192-token cutoff; final correctness uses the 768-token response. Absence of a reported early error does not certify an entirely valid reasoning trace.

Early-error classification	Final outcome	Initial	d100 seed 1
Reported early error	Correct	39	18
	Incorrect	139	248
No reported early error	Correct	560	495
	Incorrect	62	39
All outputs		800	800

The early-error prevalence increases by 11.00 percentage points, from 178/800 to 266/800. Conditional final success falls by 15.14 points among outputs with a reported early error, from 39/178 to 18/266, while it increases by 2.66 points among the remaining outputs, from 560/622 to 495/534. These are observed conditional rates in different checkpoint-induced groups. They do not establish the effect of changing the continuation checkpoint on a common erroneous prefix, and a correct final answer need not imply that an earlier erroneous step was explicitly repaired.

Early termination. The 256-token branch replay records a first EOS before token 192 in 214/800 initial Llama responses and 345, 347, and 343 adapted responses. Its corresponding Qwen counts are 195 initially and 137, 134, and 119 after adaptation. These counts include both correct and incorrect responses. The nominal cutoff can fall after termination, so an appended instruction may reopen a completed answer. Elicitation uses the replay records; budget and early-pair analyses use the separate native 768-token generations.

Common non-terminated subset. As a post-hoc sensitivity analysis, we retain a question–sample identifier only when its native first-EOS position is at least 192 in the initial state and all three adapted states. This gives 605 identifiers across 39 questions for Qwen and 365 across 38 questions for Llama. The Qwen interaction is $+5.79$ points under strict extraction, with interval $[2.55, 9.15]$, and $+9.20$ under expanded extraction, with interval $[5.34, 13.55]$. Llama gives -4.20 points, with interval $[-8.03, 0.00]$, and -0.91, with interval $[-4.60, 3.00]$, respectively. These are descriptive comparisons conditional on observed termination behavior, which can itself change after training. They do not define a causally fixed population or replace the full-sample estimates. Question resampling retains all selected identifiers for each drawn question and recomputes the response-weighted mean.

Answer retention. Under strict scoring, the answer-now branch loses 0/1/4/1 prefix successes for Llama's initial/seed1/seed2/seed3 states. Bare continuation loses 0/1/0/0. Qwen's answer-now and

bare-continuation branches lose none of their strict prefix successes. The instructed-continue branch has the larger Qwen losses reported above. A lost strict success need not be a semantically wrong answer under an alternative extraction rule.

The earlier continuation-only evaluation checked whether a correct answer appeared within the new 64-token segment, then compared it with correctness anywhere in the original prefix. Here both sides use the complete available response. Scoring the complete response changes 7,757 strict or expanded branch judgments without changing the generated text. The comparisons reported here use these full-response scores.

E Broader Budget and Transfer Results

The additional evaluations vary checkpoints, prompt rendering, training steps, and tasks to examine where the observed budget effects recur. They do not independently identify the prefix mechanism.

Rendering changes the training comparison. For initial Qwen on P, the reported 768-token score is 99.375% with chat rendering and 93.375% with plain rendering; for its seed-1 adaptation these values are 98.000% and 92.125%. At 192 tokens, the corresponding plain-rendered scores are 12.250% and 12.500%, distinct from the chat values in Table 2. Combining a chat short-budget value with a plain long-budget value would therefore mix changes in budget with changes in rendering.

The additional linear-system comparisons use R1-Distill-Qwen-1.5B (Guo et al., 2025) and Phi-3-mini-4k-instruct (Abdin et al., 2024), with chat rendering. Table 27 reports the available seed-level gains relative to each model's initial checkpoint. R1's positive long-budget gains at d200 do not demonstrate a positive-to-negative reversal. Phi-3's short-budget gains at d100 straddle zero, followed by losses at 768 tokens. Thus the consistent three-seed Llama pattern is not shared by these comparisons; no prefix or gate–content decomposition is inferred for these models.

Table 27: **Additional model comparisons do not establish a uniform budget reversal.** Chat-rendered linear-system evaluations; entries are reported seed-1/2/3 gains in percentage points, rounded as in the endpoint summaries. R1 and Phi use different training doses. A dash denotes an unreported endpoint here, not a zero gain.

Checkpoint	Steps	Gain at 192	Gain at 768
R1-Distill-Qwen-1.5B	200	—	$+8.4/+8.6/+7.5$
Phi-3-mini-4k-instruct	100	$+1.12/-0.38/+0.25$	$-4.50/-6.25/-4.63$

Native d100 trajectories. The native trajectory diagnostics use the common question set P, with IDs `alg_0050–alg_0099`, for the initial and d100 checkpoints. They describe the original budget comparison; the seed-level d200 EOS comparisons are reported separately in Appendix A.1. Initial Llama has a median output length of 245.5 tokens, compared with 208–210 across its three d100 runs. Initial Qwen has a median of 263 tokens, compared with 310–313 after d100 adaptation. These length summaries do not establish that length alone causes the accuracy changes.

Using native output lengths, initial Llama has 217 responses ended by token 192, of which 181 are correct; 418 of the 583 remaining active histories are correct at token 768. The three d100 runs have 346–349 ended responses, 269–286 ended successes, and 234–241 final successes among 451–454 active histories. Initial Qwen has 195 ended responses, 194 ended successes, and 601 final successes among 605 active histories. At d100 the corresponding ranges are 120–138 ended, 120–138 ended successes, and 646–670 final successes among 662–680 active histories.

For Equation 18, the exact ended-error counts on P are 36 initially and 74/63/77 at Llama d100; Qwen has one initially and zero in its three d100 runs. The rate uses native output length and the 192-token judgment. These are native-stream populations; Section 4 reports the separate exchange populations and their continuation outcomes.

Earlier algebra selection comparison. Consider choosing between an initial model and its 100-step adaptation, with success at 768 tokens as the evaluation objective. A rule that retains the

adaptation when its 192-token gain is positive selects all three adapted Llama checkpoints. Their losses at 768 are 10.75, 9.00, and 12.00 points. The same rule retains the initial Qwen model for all three seeds. Across these six comparisons, the average selection loss is $31.75/6 = 5.29$ points.

These decisions are evaluated on the same questions used to compute the scores. Always retaining the initial model incurs zero loss in all six comparisons because every adapted checkpoint is worse at 768. Matching that baseline would not establish a useful diagnostic. The full-response answer-now score selects the three adapted Llama checkpoints and makes only three correct decisions out of six (Table 23).

F Minimum-Length Intervention

The minimum-length comparison uses the initial and d100 seed-1 Llama checkpoints, the existing 50 questions per task, and 16 responses per question. The reported intervention prevents termination before 192 newly generated tokens and permits normal termination thereafter, under a total cap of 768 tokens. Algebra uses strict extraction and GSM8K uses the numerical extraction rule in Appendix C. The normal counts of 599/513 on algebra and 606/525 on GSM8K match the single-pass comparisons reported earlier. The adapted algebra normal and Min-192 evaluations were rerun separately.

Decoding policy. Let $q_m(a \mid h)$ denote checkpoint m's next-token distribution at history h. Before the minimum length $L = 192$, the intervention samples from

$$q_m^{\min}(a \mid h) = \begin{cases} 0, & a = \mathrm{EOS}, \\ q_m(a \mid h)/(1 - q_m(\mathrm{EOS} \mid h)), & a \neq \mathrm{EOS}, \end{cases} \tag{14}$$

and restores q_m thereafter, with the same 768-token cap. We assume $q_m(\mathrm{EOS} \mid h) < 1$ at every reachable history before L. Renormalization preserves the relative non-ending token probabilities at that history. Continuing past a stopping point can still change later histories.

Table 28: **A minimum generation length improves adapted Llama more than initial Llama.** Correct-response counts out of 800 at a total output cap of 768 tokens. Min-192 disallows termination before 192 new tokens. Changes are in percentage points. The comparison uses one adapted training seed per task.

Task	Checkpoint	Normal	Min-192	Change (pp)
Algebra	Initial	599	605	+0.750
	Adapted	513	572	+7.375
GSM8K	Initial	606	613	+0.875
	Adapted	525	577	+6.500

We subtract the initial checkpoint's net improvement from the adapted checkpoint's net improvement. On algebra, $(572 - 513) - (605 - 599) = 53$ additional correct responses reduce the normal deficit of 86. On GSM8K, $(577 - 525) - (613 - 606) = 45$ reduce the deficit of 81. Dividing by 800 gives 6.625 and 5.625 percentage points. The remaining deficits are 33 and 36 responses. These comparisons use the same decoding condition for both checkpoints; comparing adapted Min-192 against initial normal generation would answer a different model-selection question.

The minimum-length intervention and prefix exchange measure different effects. The former measures how the native checkpoint comparison responds when both checkpoints receive the minimum-length policy, $(C_t^{\min} - C_0^{\min}) - (C_t^{\mathrm{native}} - C_0^{\mathrm{native}})$. It removes EOS at the current history and thereby changes which later histories are reached. The prefix exchange holds the two-stage continuation procedure fixed and changes the prefix source. Its diagonal cells can differ from single-pass generation because continuation randomness and batch layout restart. The different decoding policies and reference scores prevent adding the two effects as parts of the fine-tuning loss.

The available evidence is aggregate correct-response counts. The execution configuration, question-level and paired response scores, generation lengths, and latency measurements are not included. The reported aggregates support accuracy comparisons, but not significance estimates or compute-normalized benefits. The algebra difference of 59 correct responses compares aggregate totals from separate runs; it does not count paired wrong-to-right transitions.

F.1 What delaying early termination can recover

Consider an intervention that only suppresses ending before m non-ending tokens, then restores the original stopping rule. Let T count the non-ending tokens generated before native termination and let $Y \in \{0, 1\}$ be the native final score (one for correct, zero for incorrect), under a common maximum output allowance. If the conditional non-ending token distribution is unchanged, its population accuracy effect satisfies

$$-\Pr(T < m, Y = 1) \;\le\; C_{\min} - C_{\text{native}} \;\le\; \Pr(T < m, Y = 0). \tag{15}$$

To see this, couple the two processes using the same non-ending token draws and stopping uniforms. They agree until a native ending before m. Every other trajectory has the same final score. Among trajectories that diverge, a native error can improve by at most one and a native success cannot improve. Reversing these roles gives the lower bound. Correctness need not be monotone during continuation. The bound concerns this decoding intervention; it does not model how SFT changes the checkpoint.

Table 29 applies the repair calculation to a fixed set of recorded native trajectories. Its candidate pool combines early-termination flags from the separate replay and native calls, including the cutoff boundary and restricting to native final errors. With every listed candidate repaired and all other outcomes unchanged, each adapted count would increase by the pool size. We compare this total with the initial model's 599 native successes.

Table 29: **The candidate early-error pool is smaller than the native deficit.** Counts out of 800 under strict algebra scoring. Replay and Native are the early-error readouts from the 256-token replay and native 768-token generation, respectively. Candidates include the union of flags at or before 192, restricted to native final errors. "All repaired" adds Candidates to Native correct. This calculation concerns the recorded native trajectories, not an independent generation run.

Seed	Native correct	Deficit	Replay	Native	Candidates	All repaired
1	513	86	72	73	74	587
2	527	72	62	62	63	590
3	503	96	74	75	77	580

Even adding all candidate repairs leaves 12, 9, and 19 fewer correct responses than the native initial-model reference. These totals concern only the recorded candidate pool; sampling uncertainty remains when Equation 15 is used to make a population claim. They do not bound an arbitrary replacement of the stopping policy, which may introduce earlier as well as later stops, or a comparison in which both checkpoints are changed. The separately rerun Min-192 evaluation has 59 more correct responses than normal decoding; these aggregate totals do not identify recovered cases or validate the fixed-trajectory repair bound.

G What the Exchange Identifies

Fix the question distribution, nominal handoff position, continuation allowance, sampling procedure, and full-response scorer. For a history h, let $K_j(y \mid q, h)$ be the experimental continuation distribution, including empty continuations when the implementation produces them. If $g(q, h, y) \in \{0, 1\}$ is correctness of the complete response, then $v_j(q, h) = \mathbb{E}_{y \sim K_j}[g(q, h, y)]$. Iterated expectation gives Equation 2. No monotonicity of correctness with length is required. In particular, terminated prefixes remain in μ_i and are handled by K_j as in the actual experiment. Treating every ended prefix as absorbing would describe a different procedure when the handoff resumes generation.

A construction with identical diagonal scores. Take one question and two possible prefix states a, b. For the first matrix in Equation 3, let μ_0 put all mass on a, μ_t all mass on b, and set $v_0(a) = v_t(a) = p$, $v_0(b) = v_t(b) = r$. Only the prefix distribution changes. For the second matrix, choose any common prefix distribution $\mu_0 = \mu_t$ and set $v_0(h) = p$, $v_t(h) = r$ on its support. Only continuation success changes. Correct and incorrect terminal outputs sampled with these probabilities realize both constructions. Thus the diagonal success probabilities alone do not determine whether the observed gain follows the prefix or continuation component. The construction motivates the off-diagonal comparisons; it does not explain the behavior of a particular checkpoint.

Interactions and reference choices. The two replacement orders in Appendix B differ in their allocation by $J = C_{tt} - C_{t0} - C_{0t} + C_{00}$. Thus a single allocated prefix effect depends on the reference or averaging convention. The main text and the exact cell tables retain all four controlled contrasts, including their unequal values when $J \neq 0$.

For algebra Llama, the smaller absolute prefix contrast exceeds the larger absolute continuation contrast by 4.125, 4.750, and 6.750 points across the three seeds. On the full GSM8K test set this margin is 6.501 points for Llama. Qwen has the reverse ordering in all three seeds: its smaller continuation contrast exceeds its larger prefix contrast by 3.393, 3.848, and 3.791 points. These statements describe the observed cells on the fixed evaluation questions, not a bound on population estimation error.

The contrasts compare distributions of complete histories, including content and termination state. Changing the source also changes the realized prefixes, so the comparison is not restricted to a common set of unfinished or incorrect responses. They also refer to two-stage generation: single-pass scores cannot replace diagonal cells without changing the experiment. These comparisons identify which generation stage contributes to the loss under the handoff procedure, not which internal reasoning operation has changed.

H Local First-Error Intervention

Selection and treatment. The local intervention uses the initial Llama checkpoint and d100 seed 1 on the existing 50-question algebra and GSM8K diagnostic collections, applying the same content-error definition on both tasks. An eligible case contains a complete, verifiable intermediate error before token 192. The first erroneous step determines the cut, and all later native text is discarded. Originally correct and incorrect final responses are both eligible. Cases without a verifiable local replacement, with an incomplete error at the cutoff, or requiring the full final answer as the replacement are excluded. The resulting algebra sets contain 146 initial-source and 234 adapted-source prefixes, compared with 178 and 266 early-error outputs in the native annotations. Each source therefore excludes 32 candidates. The GSM8K sets contain 48 and 76 eligible prefixes. Category-specific exclusion counts and question coverage are not available in the aggregate report.

In O, the incorrect step is retained. In R, only that step is replaced with a correct version, verifiable from the question and the preceding valid steps, without providing the complete final answer. In C, the incorrect step is retained and followed by a newline and `Recheck the preceding calculation, then continue.` For algebra, a deterministic repair procedure searches for at most two edits within the erroneous line, ordering candidates by edit distance, numerical change, and leftmost position. A replacement must reparse and pass the task verifier; replacements revealing the complete final answer are rejected. Both continuation checkpoints receive every selected case under all three treatments. The procedure uses temperature 0.8, top-$p = 1$, one continuation per case and treatment, and a cap of 576 new tokens. Final correctness is judged on the available prefix plus continuation with the corresponding task's existing scoring rule. These are new generations from the error location, not rescored native responses or the diagonal cells of the 192-token exchange.

Table 30: **Correct continuation counts in the local first-error experiment.** O retains the first error, R corrects it, and C requests a general check. Each pair of rows uses the same eligible cases under all three treatments. N counts response prefixes, not independent questions.

Task	Prefix source	Continuation	N	O	R	C
Algebra	Initial	Initial	146	9	102	11
	Initial	Adapted	146	8	87	7
	Adapted	Initial	234	12	136	33
	Adapted	Adapted	234	8	114	23
GSM8K	Initial	Initial	48	3	29	5
	Initial	Adapted	48	2	21	4
	Adapted	Initial	76	5	42	8
	Adapted	Adapted	76	3	27	6

Interpretation and measurement. Within each prefix source, all treatments use the same cases. Across sources, the selected errors differ, so their treatment rates do not isolate a prefix-source effect on a common population. Table 30 reports net success differences; the counts do not identify which responses were repaired or whether a model explicitly corrected itself.

Only one continuation is sampled per case and treatment, making the counts sensitive to decoding variability, especially on the smaller GSM8K subsets. Repeated responses are clustered within questions, and the aggregate counts cannot quantify question-level variation or independently validate individual edits. The correction is externally supplied, so this experiment does not evaluate an autonomous decoding method or its deployment benefit.

Table 31: **Mean newly generated tokens in the local first-error experiment.** The case collections and success counts are those of Table 30. O, R, and C share an output cap of 576, but their input lengths and actual computation are not equated. Token counts exclude the supplied prefix and intervention text; runtime and editing cost are not measured.

Task	Prefix source	Continuation	O	R	C
Algebra	Initial	Initial	139.2	139.5	198.0
	Initial	Adapted	137.4	145.1	146.6
	Adapted	Initial	156.7	163.0	212.6
	Adapted	Adapted	145.1	143.3	149.6
GSM8K	Initial	Initial	120.5	124.8	188.3
	Initial	Adapted	118.7	128.6	179.5
	Adapted	Initial	140.2	146.7	205.4
	Adapted	Adapted	138.9	151.3	198.6

I How Local Correction Changes the Overall Model Gap

For checkpoint $m \in \{0, t\}$, let E_m be its eligible error-prefix set within its 800 native responses. Use only the rows whose prefix source and continuation checkpoint are both m in Table 30. Let S_m^z be the correct count on E_m under treatment z, and let U_m be the native correct count outside E_m. Holding unselected outputs identical across treatments defines the following constructed score

$$\widetilde{C}_m^z = \frac{U_m + S_m^z}{800}, \qquad \widetilde{\Delta}^z = \widetilde{C}_t^z - \widetilde{C}_0^z. \tag{16}$$

The individual U_m values are unavailable, but cancel when changing the local treatment:

$$\widetilde{\Delta}^R - \widetilde{\Delta}^C = \frac{(S_t^R - S_t^C) - (S_0^R - S_0^C)}{800}. \tag{17}$$

This identity requires the same eligible cases and unchanged outcomes outside them across treatments. It uses the aggregate correct counts without assuming which individual responses changed from wrong to right. With $n_m = |E_m|$, each checkpoint's full-collection contrast equals its eligible fraction $n_m/800$ times its conditional difference $(S_m^R - S_m^C)/n_m$.

For algebra, the initial checkpoint's conditional difference is $91/146$, compared with $91/234$ for the adapted checkpoint. Multiplying by the respective eligible fractions gives $91/800$ for both, so Equation 17 is zero. On GSM8K, the counts are $24/800$ and $21/800$, giving $-3/800 = -0.375$ percentage points. Replacing C by O instead gives an adaptation-gain change of $+1.625$ points on algebra and -0.250 on GSM8K. The comparator therefore matters.

Equation 17 compares local treatments with the same eligibility and handoff rule, combining recorded native prefixes with newly sampled continuations. The error-location cut and 576-token continuation cap differ from uninterrupted 768-token generation, so the constructed scores are not end-to-end samples from a deployed policy. Their variation over questions and training seeds has not been quantified; equality in this construction does not establish population-level equality or predict future ordering.

Bounds on the constructed correction score. Let k_m count native successes within E_m. Then $U_m = K_m - k_m$, where K_m is the known native correct count, and $800\widetilde{C}_m^R = K_m - k_m + S_m^R$. On algebra, the initial model's 178 early-error outputs include 39 native successes, and 146 are selected. Hence $7 \le k_0 \le 39$. For the adapted checkpoint, 234 of 266 early-error outputs are selected and 18 are native successes, giving $0 \le k_t \le 18$. The constructed correction counts are therefore $701 - k_0$ and $627 - k_t$, leaving the initial checkpoint ahead by 35–85 responses. On GSM8K, the available counts give only $0 \le k_0 \le 48$ and $0 \le k_t \le 76$. The corresponding counts are $635 - k_0$ and $552 - k_t$, leaving a difference of 35–159 responses. Thus the recorded correction outcomes cannot close the initial checkpoint's lead under this substitution construction, for any compatible allocation of the unreported native outcomes. These bounds follow from the reported counts under the substitution rule; they do not estimate sampling uncertainty or constrain other correction procedures.

J Scoring and Adapter Selection

Answer extraction can reverse the training comparison. In the original 50-question GSM8K diagnostic subset, the prompts request no particular answer marker. Numerical extraction gives Llama a two-stage gain of -7.000 points. Accepting only #### or an explicit answer cue instead gives 59/800 correct initial-to-initial outputs and 254/800 adapted-to-adapted outputs, a gain of $+24.375$ points on identical text. This score requires both a cue and a correct extracted value. Qwen remains positive when an answer cue is required, but its corresponding counts fall to 2/800 and 7/800. Using only the final number reproduces Llama's numerical scores and preserves Qwen's positive gain (Appendix C). When the prompt imposes no format requirement, missing an answer cue should not be interpreted as giving a numerically wrong answer.

Changing an extraction rule leaves the generated text fixed; changing a continuation instruction does not. In a separate 64-token algebra experiment, an answer-now instruction benefits adapted Qwen more than its initial checkpoint, while Llama does not show the same interaction. Appendix D reports the full-response scores and their sensitivity to extraction.

A short snapshot can retain a harmful adapter or reject a useful one. A positive 192-token gain would select all three adapted Llama checkpoints on algebra, despite their 9.0–12.0-point losses at 768 tokens. On the original 50-question seed-1 GSM8K diagnostic comparison, the same rule correctly retains initial Llama but also retains initial Qwen, whose early gain is -0.5 points. It would forgo a 5.125-point two-stage gain, or a 4.375-point gain in the separate single-pass comparison. These are selection outcomes computed from the observed evaluations, not a test of a new learned selection method.

These examples show the cost of choosing a checkpoint at a budget different from the one at which it will be used. Prefix exchange helps locate the source of the discrepancy, but the experiments do not establish a cheaper or more accurate selection rule. No runtime saving is inferred from the ratio of output caps. Appendix E reports the earlier selection baselines and task extensions.

K Checkpoint Selection and Budget Scoring

An upper bound from early termination. Let

$$W_m^L = \Pr_m(\text{ended by } L \text{ and incorrect at } L) \tag{18}$$

be the fraction of responses that have ended incorrectly. Under a common decoding policy with only the maximum cap changed, if EOS is absorbing and ended responses retain their scores, then for any $H \ge L$,

$$C_m^H \le 1 - W_m^L. \tag{19}$$

Every ended error remains incorrect at H, irrespective of the model architecture or future continuation quality. This yields a sufficient condition for rejecting an adapted checkpoint before generating its long responses:

$$1 - W_t^L < C_0^H \quad \Longrightarrow \quad C_t^H < C_0^H. \tag{20}$$

Equation 20 suggests an evaluation sequence: measure short-budget accuracy and the fraction of ended errors, and compare the resulting upper bound with a stored long-budget score for the

initial checkpoint. This can sometimes rule out an adapted candidate before long generation. The remaining candidates must be evaluated at the operating budget. If their ranking changes, the four crossed cells show whether the difference follows prefix generation, continuation, or both. The correctness–termination table can then suggest an intervention, but does not establish the cause by itself.

Scope of the selection analysis. The selection analysis uses native initial and d100 outputs on the common question set P. The d200 EOS intervention comparisons in Section 6 address a separate question and are not used to validate the early-termination selection rule.

The cutoff catches the two models at different phases. In the per-response records, the median position at which a correct answer first appears moves from token 227 to 173 after Llama adaptation, but from 252 to 298 after Qwen adaptation. At token 192, the fraction with no correct answer yet seen falls for Llama and rises for Qwen, matching the direction of their short-budget changes. At the longer endpoint, 35.9% of Llama seed-1 adapted responses have still never produced a correct answer, compared with 25.1% initially. The sign of the short-budget training gain reflects which successes become visible before the cutoff; the sign at the longer budget also depends on how many histories succeed by that endpoint.

On the common recorded P set, Llama's W^{192} changes from 4.50% initially to 7.88–9.63% at d100, while Qwen changes from 0.125% to zero. Llama's d100 upper bounds remain above its initial long-budget score, so the bound is inconclusive and long adapted generation is required.

The common recorded P set supports six d100 checkpoint comparisons: three per family. Short-score selection favors adapted Llama in its three runs even though each loses at 768 tokens; both budgets favor initial Qwen in its three runs. The population bound remains sufficient under absorbing termination and a common response population, but the evaluated comparisons do not demonstrate a successful early rejection of an inferior checkpoint. Outcomes on these six comparisons are not an estimated error rate for new checkpoints.

Paired prefixes and independent short generation. We score 192-, 384-, and 768-token prefixes from the same long responses, so each budget comparison follows a fixed sampled trajectory. Separate generations capped at 192 tokens check the short-budget endpoint. The eight model-state collections share question and sample identifiers. Table 2 distinguishes independent short-generation scores from truncated-prefix scores. The timing analysis applies the same extractor to token prefixes and records the first position at which the extracted answer is correct. Its median is conditional on responses that produce a correct answer within the inspected stream, so a shift can reflect a different set of successful responses as well as earlier answer appearance within a response.

Intervention and exchange protocols. Appendix A.1 specifies the EOS training ablation and the gate–content policies; Appendix F specifies the inference-time minimum-length policy. Their effects use different controls and cannot be added as fractions of the adaptation deficit. Algebra handoff details are in Appendix B, and full-test GSM8K generation and extraction are in Appendix C. The task-transfer evaluations and their distinct denominators are given in Appendix A.

L From Terminal Supervision to Stopping Effects

The training signal. At a fixed history h, write $q_\theta(\mathrm{EOS} \mid h) = s_\theta(h)$ and $q_\theta(y \mid h) = [1 - s_\theta(h)]r_\theta(y \mid h)$ for a non-EOS token y. For $0 < s_\theta(h) < 1$, define $a_\theta(h) = \log[s_\theta(h)/(1 - s_\theta(h))]$. Token cross-entropy then gives

$$\ell_{\mathrm{EOS}}(h) = -\log s_\theta(h), \qquad \partial_a \ell_{\mathrm{EOS}} = s_\theta(h) - 1, \tag{21}$$

$$\ell_y(h) = -\log[1 - s_\theta(h)] - \log r_\theta(y \mid h), \qquad \partial_a \ell_y = s_\theta(h). \tag{22}$$

Terminal targets supply positive stopping labels; content targets supply both negative stopping labels and token-identity supervision. In independent output coordinates (a, r), the terminal loss has no r term. The network does not have independent parameter blocks for these coordinates, so this observation does not imply that terminal supervision leaves conditional content unchanged. Both training conditions retain the original loss denominator, so masking EOS does not reweight the remaining token losses.

Transfer to a generated history. A single small, EOS-only ordinary-SGD step at a fixed training history h is

$$\delta\theta = \eta[1 - s_\theta(h)]\nabla_\theta a_\theta(h). \tag{23}$$

For a fixed evaluation history h', a first-order expansion yields

$$\delta a_\theta(h') = \eta[1 - s_\theta(h)]\langle\nabla_\theta a_\theta(h'), \nabla_\theta a_\theta(h)\rangle + O(\eta^2), \tag{24}$$

$$\delta \log r_\theta(y \mid h') = \eta[1 - s_\theta(h)]\langle\nabla_\theta \log r_\theta(y \mid h'), \nabla_\theta a_\theta(h)\rangle + O(\eta^2). \tag{25}$$

At the training history itself, the first inner product is a squared norm and is nonnegative. At a generated history it can have either sign. Thus short demonstrated answers alone do not imply premature stopping on unresolved evaluation histories. Positive cross-history alignment would transfer the stopping signal; the second inner product describes simultaneous content change. Non-EOS losses add their own updates. These are local sensitivity identities, not a reconstruction of multi-step LoRA training or its optimizer state. The following diagnostic measures the EOS-log-odds effects; it does not measure projections onto conditional content probabilities.

Checkpoint measurements. We evaluate standard-SFT and EOS-ablated d200 seed-1 checkpoints on the same algebra histories within each family. Each family has two groups of 20 prefixes: the history immediately before an incorrect early termination, and an incorrect history still active at token 192. These 40-prefix diagnostic sets are separate from the 800-response endpoint evaluations. Let L_{EOS} and L_{content} denote the terminal-target and nonterminal-target components of the supervised loss used for the gradient diagnostic. At the standard checkpoint, define

$$D_E(h) = -\langle\nabla_\theta a_\theta(h), \nabla_\theta L_{\mathrm{EOS}}\rangle, \qquad D_C(h) = -\langle\nabla_\theta a_\theta(h), \nabla_\theta L_{\mathrm{content}}\rangle. \tag{26}$$

The two loss components use a common normalization, so their sum gives the full-loss projection. These quantities predict EOS log-odds changes per unit ordinary-SGD step size, not probability changes or actual Adam updates. Table 32 reports means and the positive/negative counts computed for each quantity. We count full-loss signs after adding the two projections for each prefix; the marginal sign counts cannot be added.

Table 32: **Terminal supervision pushes toward stopping on incorrect algebra histories.** Both families use d200 seed 1; every row contains 20 prefixes. EOS probabilities compare standard and ablated checkpoints on identical histories. Gradients are evaluated at the standard checkpoint. The last column lists positive/negative counts for D_E, D_C, and their sum, in that order. Early-end histories are measured before EOS; active histories have not terminated at token 192.

Model	History	EOS std./abl.	D_E	D_C	$D_E + D_C$	Signs: E; C; sum
Llama	Early end	0.203/0.088	$+7.6$	-5.1	$+2.5$	19/1; 2/18; 17/3
	Active at 192	0.047/0.019	$+2.9$	-2.4	$+0.5$	18/2; 3/17; 15/5
Qwen	Early end	0.169/0.097	$+5.3$	-4.4	$+0.9$	18/2; 3/17; 14/6
	Active at 192	0.036/0.022	$+1.9$	-1.6	$+0.3$	17/3; 4/16; 13/7

In both families, the early-end and active-at-192 groups have positive mean terminal-loss projections and negative mean content-loss projections. Their full-loss means are positive, with positive sums on 13–17 of 20 histories per group. The trained EOS-ablated checkpoint also has a lower mean stopping probability in every same-history comparison. These measurements show how terminal supervision can increase the tendency to stop on the sampled incorrect histories. The smaller proportion of positive net projections in Qwen's active group (13/20) shows that the stopping effect varies across histories. The diagnostic measures local gradient directions at these checkpoints; it does not reconstruct their training trajectories or test other seeds or MBPP.

Does stopping prevent recovery? We reuse the 20 prefixes preceding incorrect early termination in each family and add 20 prefixes preceding correct termination as controls. Each history is evaluated at the standard d200 seed-1 checkpoint with 16 continuations under each of two policies: normal sampling, or suppressing EOS at the first continuation step only and then restoring normal sampling. Both policies use the same total 768-token cap, including the prefix. The history labels refer to the original sampled response; resampling a continuation can change its outcome. Each row therefore contains 320 continuations per policy, rather than 320 independent prefixes.

Table 33: **A one-step stopping intervention recovers some incorrect histories.** Both models use algebra d200 seed 1. N_+ counts prefixes with $D_E + D_C > 0$; gradient entries are group means. Normal and masked columns give correct counts out of 320. The last column restricts the comparison to the N_+ prefixes, with denominator $16N_+$. All changes are percentage points.

Model	History	D_E/D_C	N_+	Normal	Masked	Change	Positive-sum subset
Llama	Incorrect end	$7.6/-5.1$	17	92	114	$+6.88$	$79 \to 98/272$; $+6.99$
	Correct end	$8.2/-4.6$	18	268	259	-2.81	$242 \to 234/288$; -2.78
Qwen	Incorrect end	$5.3/-4.4$	14	86	104	$+5.63$	$68 \to 84/224$; $+7.14$
	Correct end	$5.9/-3.7$	16	276	272	-1.25	$221 \to 218/256$; -1.17

Suppressing EOS for one step raises success on incorrect-end histories by 6.88 points in Llama and 5.63 in Qwen, while reducing it on correct-end controls by 2.81 and 1.25 points (Table 33). The incorrect histories with positive full-loss stopping projections also improve, by 6.99 and 7.14 points. Across evaluated prefixes, approximately 72% have a higher empirical continuation success rate under the intervention; ties are not counted as improvements. This is a prefix-level direction summary, not a test of significance for each prefix.

The larger mean stopping projections in correct-end controls show that stopping pressure is not specific to errors: it can preserve a correct response or prevent recovery from an incorrect one. The gradient and intervention measurements connect training signals to these local consequences on the same histories, while leaving the accumulated effect of training unresolved.

The value of continuing. Consider one controlled generation segment with a finite token cap H, fixed conditional content policy r, and alternative stopping gates s and s'. Let $u(h) \in [0, 1]$ be the utility assigned when the gate ends the segment at history h, including any prescribed EOS handling at the handoff. Ending the segment is absorbing within this comparison. At the cap, let $u_H(h)$ be the common terminal evaluator, which may differ from the EOS-ending evaluator. Let $V_s(h)$ be expected terminal utility under (r, s), with $V_s(h) = u_H(h)$ at the cap, and let

$$Q_s(h) = \sum_{y \neq \mathrm{EOS}} r(y \mid h) V_s(hy). \tag{27}$$

At an active history, $V_s(h) = s(h)u(h) + [1 - s(h)]Q_s(h)$. If $d^k_{s'}(h)$ is the probability of reaching h still active after k content tokens under (r, s'), then

$$V_{s'}(h_0) - V_s(h_0) = \sum_{k=0}^{H-1} \sum_{h} d^k_{s'}(h)[s'(h) - s(h)][u(h) - Q_s(h)]. \tag{28}$$

The probabilities $d^k_{s'}$ include the probability of remaining active; they are not conditioned on survival. The equation averages over the common initial-question distribution when there is more than one h_0.

To verify the identity, take one action under (r, s') at h, assigning value $u(h)$ to EOS and reference value $V_s(hy)$ to a content token. Subtracting $V_s(h)$ gives $[s'(h) - s(h)][u(h) - Q_s(h)]$. Sum these one-step differences along a trajectory sampled from (r, s'). The intermediate reference values cancel; termination or the cap contributes the realized utility. Taking expectations gives Equation 28. This finite-horizon identity describes the utility difference; it is not an optimization guarantee.

For full-response algebra generation, u is the terminal correctness score. For MBPP prefix generation, u is instead the expected full-response score after the fixed handoff and initial-model continuation. It is not the correctness of the prefix alone. Each gate contrast fixes the content policy and the terminal evaluator, so the identity applies separately to $B - A$ and $D - C$; their average gives G. Their occupancy distributions and reference values can differ, so neither contrast can be recovered from a mean-length statistic.

If $s' \leq s$ and $Q_s \geq u$ at every reached history, lowering the gate cannot reduce expected utility. Without these conditions, the sign depends on where stopping changes: continuation can recover an unresolved answer or spoil a correct one. Gate dominance in accuracy therefore need not mean gate dominance in parameter distance or token-distribution divergence. The four-cell contrasts measure net utility effects, while Table 32 measures local stopping pressure. Neither identifies all history-level terms in Equation 28 or reconstructs the training trajectory.

M Sensitivity to Individual Questions

The reported counts give a conservative bound on how much deleting a question could change the EOS benefit and the gate–content ordering. They concern the fixed evaluated questions and sampled responses, not uncertainty over new questions or training runs.

Let N be the number of questions and $k = 16$ the number of responses per question in each cell. Write A, B, C, D for total correct counts. The MBPP tables subtract the same initial-source reference from every cell; that reference cancels in the contrasts below. Define $T = D - A$ and $M = B - C$. Then $T/(kN)$ is the total EOS-ablation benefit, and the symmetric allocations satisfy

$$G - K = \frac{B - C}{kN} = \frac{M}{kN}, \tag{29}$$

where G, K are expressed as proportions. A question contributes at most k to either count contrast, since its cell counts lie in $[0, k]$. After deleting any common set of $r < N$ questions from all four cells, the corresponding contrasts, expressed as proportions, therefore obey

$$\text{benefit}_{-r} \geq \frac{T - kr}{k(N - r)}, \qquad (G - K)_{-r} \geq \frac{M - kr}{k(N - r)}. \tag{30}$$

The bounds hold for any choice of deleted questions, including those contributing most to the observed improvement. They require neither independent responses nor the identities of the improved questions.

Table 34: **Neither the EOS benefit nor average gate dominance can be reversed by deleting one question.** Each bound is the minimum over the three reported seeds, rounded down to three decimals. Algebra uses full-response accuracy; MBPP uses prefix value under the fixed initial continuation model. The last two columns give numbers of arbitrarily chosen questions whose removal is guaranteed to preserve the indicated positive contrast, taking the most conservative seed. These are lower bounds, not measured leave-out results or confidence intervals.

			After any one deletion (pp)		Deletions tolerated	
Task	Model/dose	N	Benefit	$G - K$	Benefit	$G > K$
Algebra	Llama/d200	50	19.897	10.586	10	6
	Qwen/d200	50	16.326	7.142	8	4
MBPP	Llama/d100	250	3.162	1.531	8	4
	Qwen/d200	250	5.773	3.263	15	9

The ordering of all four controlled effects also survives any single-question deletion: each gate contrast, $B - A$ or $D - C$, remains greater than each content contrast, $C - A$ or $D - B$. The four inequalities reduce to $M > 0$, $U = 2B - A - D > 0$, and $V = A + D - 2C > 0$. A question contributes at most k to M and $2k$ to U or V. In every row of Tables 6 and 11, $M - k$, $U - 2k$, and $V - 2k$ are all positive.

Thus no single question can account for either the benefit or the full gate-over-content ordering in any of the twelve model–task–seed comparisons. The bounds do not recover the distribution of effects across questions, establish that most questions improve, or quantify out-of-sample uncertainty.

N Further Connections to Adaptation and Generation

What changes during adaptation. Instruction-following systems combine supervised demonstrations with other stages, including learning from human feedback (Ouyang et al., 2022). The outcome also depends on the adaptation data: instruction-tuning studies vary the number and diversity of tasks (Chung et al., 2024), use small curated sets (Zhou et al., 2023), or examine mixtures and instruction formats (Longpre et al., 2023). Transfer-learning work shows that fine-tuning can distort pretrained features and underperform out of distribution (Kumar et al., 2022). We hold the adaptation examples fixed and ask why the same resulting checkpoint can improve at one output limit and deteriorate at another. An endpoint score alone cannot distinguish better partial solutions from better continuation.

Parameter-efficient updates. Adapter comparisons study how the choice of trainable modules affects language-model adaptation (Hu et al., 2023). Comparisons of LoRA with full fine-tuning also find differences in learning and forgetting (Biderman et al., 2024a), and in the structure of the learned parameter updates (Shuttleworth et al., 2025). Restricting the trainable parameters does not fix the distribution of generated histories. Our prefix and gate exchanges measure consequences within the tested LoRA setting. They do not compare adaptation methods or establish the same stage ordering under full-parameter fine-tuning.

Reasoning traces and inference resources. Chain-of-thought prompting elicits intermediate steps (Wei et al., 2022), while self-consistency samples multiple reasoning paths and aggregates their answers (Wang et al., 2023). Training on generated rationales, as in STaR (Zelikman et al., 2022), changes the model before evaluation; process supervision trains reward models with feedback on intermediate reasoning steps (Lightman et al., 2024). We compare checkpoint gains at a fixed sample count, and distinguish the timing of correct answers from the probability of subsequently recovering an incorrect prefix.

Training histories and generated histories. Scheduled sampling mixes generated and observed inputs during sequence training (Bengio et al., 2015); sequence-level training optimizes a criterion on complete generated outputs (Ranzato et al., 2016). DAgger addresses the dependence of later inputs on earlier decisions through data aggregation (Ross et al., 2011). These methods address distribution shift along a trajectory during training. Our prefix exchange diagnoses already trained checkpoints by comparing their generated histories under each continuation model.

Termination and decoding. The sequence selected by a decoder can differ from a typical sample of the model. Neural text degeneration motivates nucleus sampling (Holtzman et al., 2020); consistency analyses show that incomplete decoding can change termination guarantees (Welleck et al., 2020). In machine translation, exhaustive search reveals a model preference for excessively short or empty outputs (Stahlberg & Byrne, 2019), and beam-search analyses study uniform information density as a decoding objective (Meister et al., 2020). These findings separate properties of a learned sequence distribution from properties of the decoding rule. Our EOS factorization also separates stopping from the conditional distribution over continuing tokens, but the algebra and code experiments use sampled outputs rather than beam-search translations.

Feedback and correction. An instruction to reconsider an answer is a different intervention from providing the location or content of an error. Evidence that intrinsic self-correction can fail (Huang et al., 2024) should not be generalized to methods with different feedback. Self-Refine iterates feedback and revision (Madaan et al., 2023); Reflexion uses verbal feedback and memory across attempts (Shinn et al., 2023). Our local-edit comparison supplies a verifiable replacement for one erroneous step and holds the continuation protocol fixed. It tests whether that edit closes the observed adaptation gap, not whether a complete self-correction system can be trained.

Evaluation conditions. HELM compares language models across scenarios and multiple metrics (Liang et al., 2023), while prompt-format studies show sensitivity to semantically incidental presentation choices (Sclar et al., 2024). Reproducible evaluation also depends on details of prompting, generation, and scoring (Biderman et al., 2024b). These considerations explain why we keep the prompt, extraction rule, and sampled population fixed within each budget comparison. A change in chat rendering, answer extraction, or candidate selection answers a different question from changing the output limit. The appendix reports those sensitivity analyses separately from the prefix and stopping interventions.