SELECTIVE COMPUTATION DECLINES AFTER SHORT-ANSWER CONTINUAL FINE-TUNING: ELICITATION AND TARGET FORMAT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual fine-tuning aims to add skills while preserving existing capabilities, yet a lower benchmark score can reflect either parameter changes or a changed response protocol. We study this distinction by adapting Qwen2.5-3B on fourteen short-answer classification tasks and evaluating a held-out suite of multi-step computation and retrieval questions. Computation accuracy falls by 26.3 percentage points, whereas retrieval falls by only 1.3 points; the same ordering holds across four additional model configurations and under answer-span scoring, although the magnitude depends on the score rule. The adapted checkpoint also produces much shorter responses. Forced additional generation alone does not improve measured computation accuracy, while applying the same answer-only instruction and three-token cap to the unadapted checkpoint reproduces the selective pattern without changing weights. A protocol selected on a development split raises Qwen2.5-3B test computation from 64.7% to 81.0% under primary scoring (and from 57.0% to 71.0% under answer-span scoring), without updating the checkpoint. Across backbones, this protocol reduces the base-relative computation decline by 50.5%–70.0%. Placing the matched rationale before the answer reduces the computation decline to 3.3–4.8 points, while retrieval remains within 2.5 points of base. These controlled comparisons show that measured computation decline is sensitive to elicitation and target format.

1 INTRODUCTION

Continual fine-tuning aims to add new skills while preserving a model's existing capabilities (Shi et al., 2025; De Lange et al., 2022; van de Ven & Tolias, 2019). Recent continual-instruction benchmarks and methods study this problem with replay, instruction-aware sampling, or modular experts (Zhang et al., 2023; Wang et al., 2023; 2024b; He et al., 2024; Wang et al., 2025). Classical continual-learning evaluations typically compare performance before and after a task sequence and summarize the change as transfer or forgetting (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018; De Lange et al., 2022). For language models, however, a benchmark score reflects both the checkpoint and the protocol used to elicit and score a response. Adaptation can change the preferred response format: the checkpoint may produce a short answer even when the evaluation expects intermediate work. The score can therefore conflate what the model produces with how the protocol asks it to respond.

We study whether the measured decline depends on prompting, output constraints, and target format, and whether these dependencies differ between computation and retrieval.

Figure 1 summarizes the matched comparisons. We adapt a pretrained model on fourteen short-answer classification tasks whose targets contain a single class-label string and no written rationale. We then evaluate the unadapted and adapted checkpoints on the same frozen suite of 200 questions: 100 require combining intermediate quantities (*computation*), whereas 100 can be answered by recalling a fact or association (*retrieval*). Response interventions change prompts or generation constraints while keeping the weights fixed. A matched target-format comparison keeps the task stream and rationale content fixed while changing answer position. The design separates category-specific score changes from response-condition effects.

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

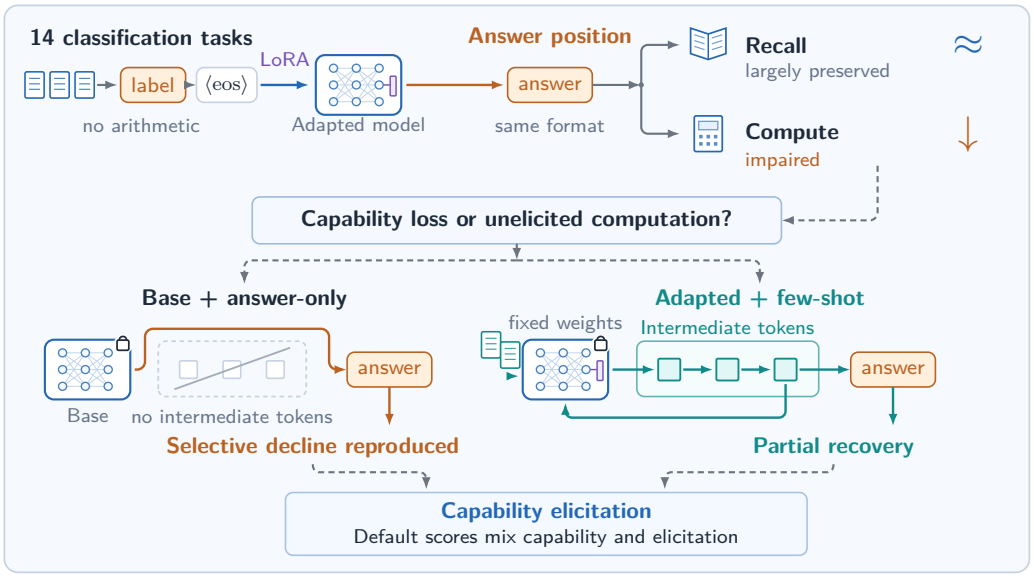


Figure 1: **A lower score after short-answer fine-tuning need not indicate lost capability.** Label-only adaptation lowers Qwen2.5-3B computation by 26.3 points but retrieval by 1.3. A base-selected prompt, an answer-only cap on the unadapted checkpoint, and rationale-first targets respectively raise, reproduce, and reduce the measured decline.

Our analysis builds on several lines of work. Fine-tuning can specialize a model to the output format of its training data and reduce generalization (Wang et al., 2024c; Zhang & Wu, 2024; Zheng et al., 2025). Continual-learning methods primarily modify training to improve retention (Wang et al., 2024b; He et al., 2024; Wang et al., 2025). Evaluation is also format-sensitive: meaning-preserving prompt changes can shift accuracy, and first-token accuracy can diverge from correctness of the generated answer (Sclar et al., 2024; Wang et al., 2024a). Free-form evaluation can conflate correctness with verbosity (Adlakha et al., 2024; Li et al., 2024). Studies of chain-of-thought fine-tuning and distillation further show that rationale supervision and target order affect reasoning behavior (Lobo et al., 2025; Wadhwa et al., 2024; Chen et al., 2025); capability-elicitation and output-format studies show that a default protocol can underestimate performance (Greenblatt et al., 2024; van der Weij et al., 2024; Singhal et al., 2023; Gan et al., 2026; Deng et al., 2026; Ravikumar et al., 2026). We combine these perspectives by varying elicitation at a fixed checkpoint and answer position at matched rationale content.

The primary setting is Qwen2.5-3B with LoRA adaptation (Qwen Team, 2024; Hu et al., 2022; Biderman et al., 2024); we repeat the main comparison across four additional model configurations. Under the default protocol, computation accuracy on Qwen2.5-3B falls from 91.0% to 64.7% (a 26.3-point decline), whereas retrieval falls from 99.0% to 97.7% (a 1.3-point decline). The same ordering holds across the additional configurations and under stricter answer-span scoring (Tables 1 and 5). Thus, the main effect is category-dependent rather than a uniform decline in accuracy.

The two categories show a similar shift in response style. Adapted answers become much shorter and usually stop immediately after the answer, while retrieval remains close to the base accuracy. This contrast motivates three response-condition tests. Allowing the adapted model to generate for longer does not improve measured computation accuracy. A 2×2 control on Qwen2.5-3B shows that an answer-only instruction and a three-token cap each lower computation accuracy on the unadapted checkpoint, while their combination reaches 65.0%, essentially the adapted default of 64.7%, without changing the weights. Finally, a few-shot protocol selected on a development split raises computation accuracy for the adapted Qwen2.5-3B checkpoint from 64.7% to 81.0% on a held-out test split (16.3 points), with the same qualitative pattern under answer-span scoring. The cross-backbone diagnostic reports a 50.5%–70.0% reduction in the base-relative computation decline (Table 1).

We next vary the training target while keeping the task stream and rationale content matched. Placing the rationale before the answer reduces the computation decline to 3.3–4.8 points on each backbone and reduces the base-relative label-only computation decline by 82.4–87.5%, while retrieval remains within 2.5 points of base. The tendency to stop at a complete answer remains similar to label-only training (95.5% versus 96.4% on Qwen2.5-3B). The comparison links answer position to the measured decline through a matched target-format intervention.

We report both the released gold-answer score and a stricter answer-span score, interpreting each intervention under its stated response conditions.

Contributions.

- We document a selective post-adaptation decline: short-answer fine-tuning sharply reduces multi-step computation while leaving retrieval nearly intact across five model configurations and two scoring rules.
- We show that the measured decline is sensitive to response conditions: on the unadapted checkpoint, an answer-only instruction and a three-token cap each lower computation accuracy, and their combination reproduces the adapted score; a base-selected protocol reduces much of the measured decline without changing weights.
- We identify training-target format as a second factor: a matched rationale-first comparison leaves a much smaller residual computation decline, whereas randomized answer position removes much of this benefit.

2 Evaluation Framework

Our central comparison holds the checkpoint fixed while varying how an answer is requested. This separates score changes caused by adaptation from those caused by elicitation. We define the comparison and then describe the checkpoints, held-out suite, scoring rules, and matched interventions.

Let M_0 denote the unadapted checkpoint and let

$$M_T^{(\tau)} = \mathrm{Adapt}(M_0; \mathcal{D}_{1:T}, \tau)$$

denote a checkpoint adapted on task stream $\mathcal{D}_{1:T}$ with training-target format τ. The primary stream has $T = 14$ tasks and label-only targets. We evaluate three adaptation seeds separately and report their mean.

Let $\mathcal{S} = \mathcal{S}_{\mathrm{comp}} \cup \mathcal{S}_{\mathrm{retr}}$ be the frozen held-out suite, with 100 items in each subset. An evaluation protocol π fixes the prompt, output instruction, decoding constraints, and scoring rule. If $h_\pi(M, x)$ is the response from M on item x, define

$$A_c(M, \pi) = \frac{1}{|\mathcal{S}_c|} \sum_{(x,y)\in\mathcal{S}_c} g_\pi(h_\pi(M, x), y), \qquad c \in \{\mathrm{comp}, \mathrm{retr}\},$$

where g_π is the binary correctness rule. Thus A_c is the accuracy for category c. With π_0 denoting the default protocol, the two changes we report are

$$\Delta_c = A_c(M_T^{(\tau)}, \pi_0) - A_c(M_0, \pi_0), \qquad R_c(\pi') = A_c(M_T^{(\tau)}, \pi') - A_c(M_T^{(\tau)}, \pi_0).$$

The first quantity measures the score change after adaptation; the second measures the change when the adapted checkpoint is held fixed and only the evaluation protocol changes. We test target-format effects by comparing Δ_c across matched values of τ.

Models and adaptation. The primary experiment adapts Qwen2.5-3B sequentially on fourteen short-answer classification tasks with LoRA (Qwen Team, 2024; Hu et al., 2022). Each training example uses a single class-label target and contains no arithmetic solution or written rationale. We repeat the comparison across model families and scales. The task list, LoRA configuration, and deterministic decoding settings are given in Appendix M.

Held-out suite and scoring. The computation subset requires combining intermediate quantities; the retrieval subset can be answered by producing a fact or association. We assign each item to a

Table 1: **Selective computation decline across five backbones.** Accuracy (%) on the frozen 200-item suite; entries after *Base* are positive declines (pp) for label-only adaptation, base-selected prompting, and rationale-first training. Label-only, prompted, and Qwen rationale-first rows average three seeds; auxiliary rationale-first rows use one seed.

	Computation (% / pp)				Retrieval (% / pp)			
Backbone	Base	Label	Prompt	R-first	Base	Label	Prompt	R-first
Phi-3-Mini (3.8B)	89.0	4027.3	1513.5	74.8	98.0	41.3	41.5	41.5
Qwen2.5 (3B)	91.0	4026.3	1510.0	43.3	99.0	41.3	73.0	41.5
Llama-3.1-Instruct (8B)	83.3	2721.6	158.3	43.8	98.7	41.7	43.4	42.5
Qwen2.5 (7B)	92.3	4027.3	159.2	43.8	99.0	41.3	42.2	41.2
Qwen3 (14B)	94.0	4026.7	158.0	43.9	99.3	41.3	41.8	41.2
Range	21.6–27.3 decline				1.3–1.7 decline			

category before evaluating either checkpoint. We report the released gold-answer score and a stricter answer-span score, and record response length and termination alongside accuracy. The two scores expose the extent to which semantic correctness and textual form diverge (Adlakha et al., 2024; Li et al., 2024; Wang et al., 2024a). The complete item list and scoring definitions are given in Appendix A–B.

Matched comparisons. The main comparison estimates Δ_c under the default protocol. We then change one response condition at a time: we allow the adapted checkpoint to continue generating, impose an answer-only instruction and a three-token cap on the unadapted checkpoint, and apply a prompting protocol selected using only unadapted-model behavior. These tests leave the adapted weights unchanged. Finally, we move the answer within the training target while matching the task stream and rationale content to test whether target format changes later response behavior.

Aggregation and statistical interpretation. Unless stated otherwise, reported accuracies for the primary Qwen2.5-3B arms are means over three independent adaptation seeds; the four auxiliary rationale-first rows use one seed per backbone. Item identities remain fixed across checkpoints, and paired item counts are reported for interventions relative to default decoding. We treat pooled item–seed values as descriptive and use paired item-level comparisons for intervention diagnostics. The released gold-answer rule is primary; the answer-span rule is a prespecified sensitivity analysis because the former can credit a long response for mentioning a gold string.

The base-selected protocol $\pi_\star$ is chosen from three candidates on a development split using the base model's computation score, then evaluated once on a held-out test split. The answer-only control tests format sensitivity, minimum-length forcing tests decoder continuation, and matched target-format arms test the training intervention. Thus, Δ_c measures checkpoint change under a fixed protocol, whereas $R_c(\pi')$ measures protocol sensitivity at a fixed checkpoint.

3 Selective Post-Adaptation Decline

The main result is a selective computation decline: under the same evaluation protocol, adaptation lowers computation accuracy much more than retrieval accuracy. We first establish this contrast across model configurations and scoring rules.

Across all five backbones, both subsets have high baseline accuracy (83.3–94.0% for computation and 98.0–99.3% for retrieval). On a separately constructed difficulty-matched diagnostic (distinct from the 200-item suite in Table 1), Qwen2.5-3B achieves 91.0% (273/300) on computation and 91.7% (275/300) on retrieval. The 0.7-point gap indicates similar observed baseline accuracy on that diagnostic. After label-only adaptation, computation falls by 21.6–27.3 points, whereas retrieval falls by only 1.3–1.7 points. Answer-span scoring preserves the same ordering (12.0–18.3 pp for computation versus 1.0–1.3 for retrieval; Appendix B, Table 5). For Qwen2.5-3B, the 1.3-point retrieval decrease comprises six item–seed losses and two gains out of 300.

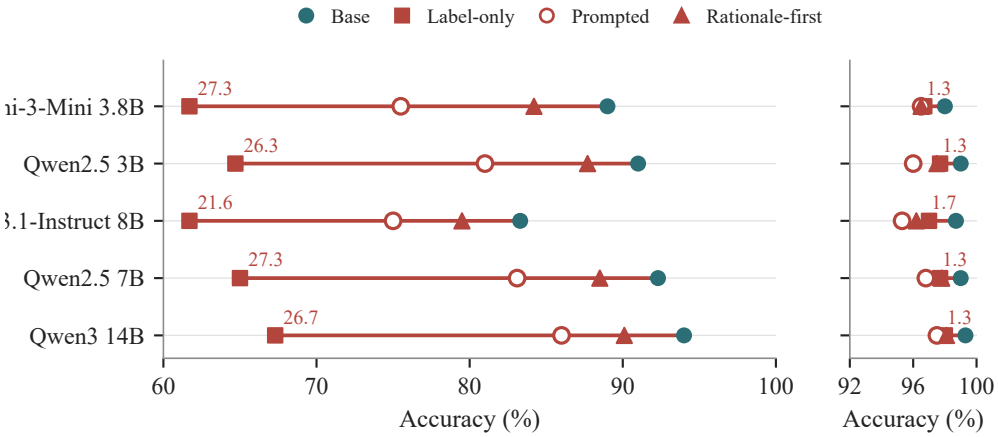


Figure 2: **The selective decline is consistent across five backbones.** Segments connect unadapted and label-only checkpoints; open circles and triangles show base-selected prompting and rationale-first training.

Table 2: **Response-condition interventions across five backbones.** Accuracy and runaway-digit rate on the frozen 200-item suite. *Default* is unforced label-only adaptation; *MinLen24* forces 24 tokens; *Answer-cap* applies a final-answer cue and three-token cap to the unadapted model; *Prompted* is base-selected. The final row gives declines relative to base.

Backbone	Base Comp	Adapted / controlled Comp (%)				Runaway
	(unadap.)	Default	MinLen24	Answer-cap (base)	Prompted	(MinLen24)
Phi-3-Mini (3.8B)	89.0	4061.7	4058.2	4063.5	1575.5	12.4%
Qwen2.5 (3B)	91.0	4064.7	4060.3	4065.0	1581.0	10.7%
Llama-3.1-Instruct (8B)	83.3	2761.7	2757.5	2762.1	1575.0	14.1%
Qwen2.5 (7B)	92.3	4065.0	4061.8	4066.2	1583.1	9.8%
Qwen3 (14B)	94.0	4067.3	4063.5	4068.1	1586.0	8.5%
Decline vs. base (pp)	0.0	21.6–27.3	25.5–30.5	21.2–25.9	8.0–13.5	—

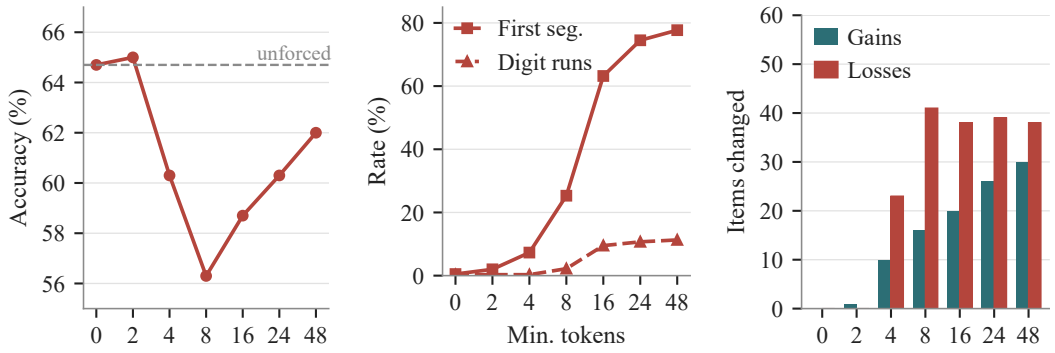
The interventions reduce part of this computation decline while leaving retrieval close to baseline. Applying the base-selected prompting protocol $\pi_\star$ (Section 4) to an adapted checkpoint without updating its weights reduces the base-relative computation decline by 50.5%–70.0%. Rationale-first fine-tuning (Section 5) changes only the training-target format and reduces the same computation decline by 82.4%–87.5%, leaving a residual decline of 3.3–4.8 points. Retrieval remains within 3.4 points of base under every reported condition and backbone.

Seed-level values remain stable across the three Qwen2.5-3B runs: label-only computation is 65/64/65%, answer-first is 59/58/59%, and rationale-first is 88/87/88%. The answer-span regrade preserves the same computation–retrieval ordering.

4 Response-Condition Interventions

The score contrast is accompanied by a large change in response style. Across all five backbones, mean response length falls from 22.4–31.5 words to 1.1–1.5 words, and $\geq$ 99.5% of adapted responses contain at most three words. Retrieval nevertheless remains within 1.3–1.7 points of base. Table 2 compares matched response conditions; Figure 3 provides the continuous minimum-generation-length sweep for Qwen2.5-3B.

Minimum-generation-length forcing does not improve measured computation accuracy. It changes when the decoder may emit an end-of-sequence token while keeping the prompt, questions, and model



(a) Computation accuracy. (b) Response degeneration. (c) Item-level changes.

Figure 3: **Extending generation does not improve computation and degrades output.** Minimum-length sweep on adapted Qwen2.5-3B (three seeds, 200-item suite). Longer generation does not sustain gains (65.0% at dose 2 versus 64.7% unforced), while contamination rises 0.5% → 77.7% and runaway digits 0.3% → 11.3%; losses exceed gains above dose 2 (Appendix I).

parameters fixed. Across all five backbones, MinLen24 lowers adapted computation accuracy relative to Default and increases runaway digit runs from $\leq 0.4\%$ to 8.5–14.1%. For Qwen2.5-3B, the full 0–48-token sweep reports accuracy, response degeneration, and item-level gains and losses relative to unforced decoding. Above dose 2, item-level losses outnumber gains and longer continuations increasingly contaminate the response before the first sentence boundary. This intervention tests continuation after the learned boundary, rather than trained test-time scaling. Per-dose values are in Appendix I (Table 12).

The combined response condition reproduces the selective pattern. On Qwen2.5-3B, a 2×2 control separates the two components: computation is 89.3% with a three-token cap alone, 74.7% with an answer-only instruction alone, and 65.0% when both are applied, compared with 64.7% for the adapted default (all values pooled over 300 item–seed evaluations; Table 8). Retrieval remains 98.3–98.7% in the three base controls. Both conditions lower computation accuracy, and their combination reproduces the adapted score without changing weights. Across the five backbones, the same answer-only-plus-cap condition yields 62.1–68.1% computation ($AnsCap_{\text{base}}$ in Table 2) and 1.0–1.4 mean words.

The two-by-two control also clarifies the minimum-length result. Capping the unadapted model alone costs 1.7 points, whereas the answer-only instruction costs 16.3 points and their combination costs 26.0 points relative to the base default. On the adapted checkpoint, granting more tokens changes the termination location and increases contamination, but does not reproduce the answer-plus-derivation behavior of the base model. In contrast, selecting prompts on development items raises test computation from 64.7% to 81.0% while retrieval moves from 97.7% to 96.0%.

The adapted responses do not collapse to a single stream label in this probe. The adapted model assigns a stream label to none of the 50 out-of-scope questions under any tested decoding condition or seed. Among 36 near-scope questions, it emits eleven distinct labels, and its most frequent label covers only 8 items (Appendix A).

Prompt selection improves computation without updating the adapted checkpoint. We select one of three candidate protocols on a development split using unadapted-model computation accuracy, then fix the selected few-shot examples for both models (Appendix P). On Qwen2.5-3B, the selected protocol raises adapted computation from 64.0% to 78.7% on the 150-item development split and from 64.7% to 81.0% on the 300-item test split; under answer-span scoring the corresponding changes are 56.0 → 69.3% and 57.0 → 71.0%. Retrieval changes from 97.3 → 96.0% (dev) and 97.7 → 96.0% (test) under the primary rule. The test-set increase is 16.3 points, reducing the base-relative computation decline from 26.3 to 10.0 points (62.0%). Selection uses only development items, so the test gain measures protocol transfer. The cross-backbone extension in Table 1 reports the corresponding decline reduction. An unconditional step-by-step scaffold performs poorly on computation and also harms retrieval (Appendix L).

Table 3: **Rationale-first targets are associated with a smaller measured computation decline across five backbones.** Pooled three-seed Qwen2.5-3B results (100 computation items per seed) and one-seed rationale-first replications on four backbones. Declines are positive relative to base; rationale arms use matched content and lengths (42 versus 43 tokens).

Backbone	Target format	Len. (tok.)	$P(\text{EOS})$ (%)	Comp. decline (pp)	Retr. decline (pp)
Qwen2.5 (3B)	(base)	–	26.9%	0.0	0.0
Qwen2.5 (3B)	Label only	1	96.4%	4026.3	1.3
Qwen2.5 (3B)	Answer → rationale	42	0.2%	4032.3	1.9
Qwen2.5 (3B)	Rationale → answer	43	95.5%	**43.3**	1.7
Phi-3-Mini (3.8B)	Label only	1	94.0%	4027.3	1.3
Phi-3-Mini (3.8B)	Rationale → answer	43	90.0%	**74.8**	1.5
Qwen2.5 (7B)	Label only	1	95.0%	4027.3	1.3
Qwen2.5 (7B)	Rationale → answer	43	91.0%	**43.8**	1.2
Llama-3.1-Instruct (8B)	Label only	1	91.1%	2721.6	1.7
Llama-3.1-Instruct (8B)	Rationale → answer	43	90.0%	**43.8**	2.5
Qwen3 (14B)	Label only	1	96.0%	4026.7	1.3
Qwen3 (14B)	Rationale → answer	43	92.0%	**43.9**	1.2

Cross-backbone transfer of protocol selection. We repeat the same development/test selection on Llama-3.1-8B, choosing the protocol from unadapted-model computation accuracy on the development split and fixing it for the test evaluation. On the held-out test split, the selected protocol raises adapted computation from 61.7% to 75.0% under primary scoring and from 54.3% to 66.0% under answer-span scoring; retrieval changes from 97.0% to 95.3%. The development gains are 14.7 and 12.6 points under the two computation rules, compared with 13.3 and 11.7 points on test (Table 11). The similar development and test changes indicate transfer of the selected protocol across this backbone.

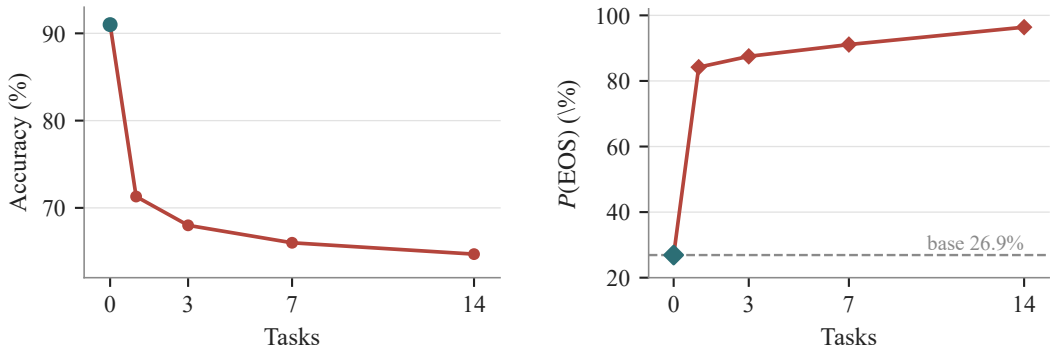
5 Answer Position in the Training Target

The response-condition tests show that adaptation changes how computation is expressed. We next test whether the training target creates this sensitivity. The label-only stream supervises a single class label; in the rationale arms, we place the same label before or after a matched rationale. We first measure termination at the answer boundary, then compare these matched targets.

Termination shifts to the answer boundary. Adaptation shifts termination toward the answer boundary rather than uniformly suppressing generation. We follow teacher-forced continuations on the 86-question diagnostic set, including the 50 out-of-scope and 36 near-scope questions. On the 26 objective out-of-scope questions, we also measure stopping immediately after the prompt and at a fixed complete-answer boundary. Adaptation lowers termination before an answer ($1.2\% \rightarrow 0.2\%$) but raises it at the complete-answer boundary ($26.9\% \rightarrow 96.4\%$). The before-answer comparison has the same direction on 25 of 26 questions ($p = 8.1 \times 10^{-7}$), and token-level measurements show the same position dependence (Appendix F, Figure 5). A general reluctance to generate would increase termination at both positions; instead, the checkpoint is less likely to stop before an answer and more likely to stop after producing one. The boundary reversal also appears on Llama-3.1-8B-Instruct ($0.8\% \rightarrow 91.1\%$).

Moving the rationale before the answer reduces the measured computation decline. The residual computation decline is smaller when the rationale precedes the answer, despite closely matched target lengths. We train three arms on the same task stream: label-only targets, an answer followed by a rationale, and a rationale followed by the answer. The rationale arms share content and have closely matched target lengths (median 42 versus 43 tokens; Appendix N). Table 3 reports termination and computation/retrieval accuracy for Qwen2.5-3B, together with rationale-first replications on four additional configurations.

On Qwen2.5-3B, placing the rationale before the answer reduces the computation decline to 3.3 points, compared with 26.3 points for label-only training and 32.3 points when the answer comes



(a) Computation accuracy.

(b) Boundary termination.

Figure 4: **Selective decline and boundary termination emerge early.** Synchronized three-seed checkpoints: computation is $91.0 \rightarrow 71.3 \rightarrow 64.7\%$ and boundary $P(\text{EOS})$ is $26.9 \rightarrow 84.2 \rightarrow 96.4\%$ after 0, 1, and 14 tasks ($N = 300$ item–seed evaluations; 26 probes).

first. Since the rationale arms match in content and target length, this contrast associates the smaller decline with answer order. Termination for the rationale-first arm remains close to label-only training (95.5% versus 96.4%); its seed-level range is summarized in Appendix N.

The same pattern appears on the four additional backbones: computation declines by 3.8 (Qwen2.5-7B), 3.8 (Llama-3.1-8B-Instruct), 3.9 (Qwen3-14B), and 4.8 points (Phi-3-Mini), a reduction of 17.8–23.5 points relative to label-only training on each backbone (Table 1, Rationale-first column). Retrieval remains within 2.5 points of base. These measurements link the smaller decline to target order and the associated answer boundary.

The two rationale orders are asymmetric: putting the answer before the rationale is worse than label-only training (58.7% versus 64.7% pooled computation), whereas putting the rationale first raises computation accuracy to 87.7%. Related distillation work also finds that rationale placement relative to labels can change outcomes, but its setting adds teacher-generated reasoning rather than continuing a continual-learning stream (Wadhwa et al., 2024; Chen et al., 2025). The asymmetry is consistent with a target-order effect on the learned response boundary. Target order also changes token-level supervision and optimization dynamics, so the result identifies answer position as a factor associated with the response policy learned during adaptation.

Randomized answer position reduces the rationale-first benefit. To separate target length and rationale exposure from a consistently placed answer, we add an auxiliary randomized-position arm. Its targets use the same rationale and answer token inventory and the same 43-token length as the rationale-first arm, but place the answer in a prefix, infix, or suffix position with equal probability across examples. The arm reaches 69.3% computation accuracy ($208/300$), a 21.7-point decline, with $P(\text{EOS}) = 68.5\%$ at the answer boundary and 9.4 mean words (Table 17). This lies above label-only training (64.7%) but well below rationale-first training (87.7%), and its termination and response length are intermediate. The comparison associates the rationale-first gain with a consistently answer-last target; because position randomization also changes token-level supervision, it does not isolate answer position as an independent causal factor.

6 Early Emergence, Depth, and Target Style

We examine when the selective decline emerges and how it varies with adaptation method and target style.

The decline is already present after one task. We evaluate computation accuracy and boundary termination at synchronized checkpoints on the same held-out items, pooling three seeds.

Table 4: **Additional scope checks.** Accuracy declines (pp) relative to unadapted checkpoints; the full-rank row holds the stream fixed, whereas the auxiliary rows vary stream and target style.

Adaptation condition	Target style	Comp. decline	Retr. decline
Full-rank fine-tuning (aux.)	Short label	24.5	1.5
Extractive question answering	Short answer span	18.5	2.1
Summarization	Long target	2.0	1.5

Adaptation method and target style. In these auxiliary checks, the decline also appears under one full-rank run. The two pilot conditions differ in both task stream and target style: extractive question answering preserves the computation–retrieval contrast, whereas long-target summarization does not. Full-rank fine-tuning on the same stream produces a 24.5-point computation decline and a 1.5-point retrieval decline, close to the LoRA declines of 26.3 and 1.3 points (Table 4).

For the Qwen2.5-3B label-only arm, seed-level computation spans 62.0–68.0% and retrieval 96.0–99.0%; the paired computation counts are 88 item-level losses versus 9 gains. On the fourteen training tasks, average accuracy changes by 0.5 points for label-only and 0.2 points for rationale-first Qwen2.5-3B, and by 0.5 and 0.2 points for Llama-3.1-8B (Appendix N).

The synchronized trajectory places the computation decline near the beginning of the stream: after one task, computation is already 71.3% while retrieval remains 97.3%. Later checkpoints deepen the same boundary shift, and full-rank training reproduces the separation in one auxiliary run. The trajectory and auxiliary checks are consistent with an interaction between continual adaptation and the response policy induced by the target.

Response protocol is therefore part of the retention measurement. Default decoding alone can fold a stopping-boundary shift into apparent capability loss, whereas a matched protocol control separates score sensitivity from checkpoint change. In this setting, target order, rationale content, and scoring rule are measurable factors; the tested rationale-first target is an intervention, not a universal prescription. The effect emerges early, persists under the tested full-rank run, and varies across the task/target pilots.

7 Conclusion and Limitations

In this controlled short-answer setting, fine-tuning reduces computation accuracy more than retrieval accuracy (26.3 versus 1.3 points on Qwen2.5-3B), with the same ordering across four additional backbones and under answer-span scoring. Forced additional generation does not improve measured computation accuracy, whereas a base-selected prompt applied to the frozen adapted checkpoint reduces the measured decline by 50.5%–70.0%. Rationale-first training leaves a 3.3–4.8-point computation decline while retrieval remains near base; randomizing answer position removes much of this benefit. The results show that continual fine-tuning can change how a capability is expressed, with answer position in the training target contributing to the measured response policy. The study covers one hand-constructed suite, one fourteen-task stream, finite protocol choices, teacher-forced probes, and limited auxiliary seeds.

AI use statement

Generative AI tools were used for language editing, document organization, LaTeX assistance, and technical consistency and proof checks. The authors independently reviewed and verified all AI-assisted material and take responsibility for the paper's arguments, claims, proofs, citations, and final presentation.

Reproducibility statement

The paper and appendices specify the held-out suite, task stream, prompts, decoding conditions, scoring rules, and paired analyses. The supplement will provide the frozen items and prompts, configuration files, item-level records, and scripts that regenerate the reported tables and figures.

Ethics Statement

This work uses publicly available NLP benchmarks and open-weight checkpoints. We do not collect new human data, recruit participants, or conduct human or animal experiments. We report scoring sensitivity and response-format limitations to reduce misuse and over-interpretation.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. In *Transactions of the Association for Computational Linguistics*, volume 12, pp. 681–699, 2024. doi: 10.1162/tacl_a_00667. URL https://aclanthology.org/2024.tacl-1.38/.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15094–15119, 2025. doi: 10.18653/v1/2025.findings-acl.782. URL https://aclanthology.org/2025.findings-acl.782/.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ale Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

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/.

Chengguang Gan, Hanjun Wei, Yunhao Liang, Qinghao Zhang, Shiwen Ni, and Zhixi Cai. How output format confounds data quality and capability in instruction tuning. *arXiv preprint arXiv:2609.02015*, 2026.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024.

Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. SEEKR: Selective attention-guided knowledge retention for continual learning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3254–3266, 2024. doi: 10.18653/v1/2024.emnlp-main.190. URL https://aclanthology.org/2024.emnlp-main.190/.

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 (ICLR)*, 2022.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Boyd-Graber. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9373–9398, 2024. doi: 10.18653/v1/2024.findings-emnlp.548. URL https://aclanthology.org/2024.findings-emnlp.548/.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*

Long Papers), pp. 11679–11698, 2025. doi: 10.18653/v1/2025.naacl-long.584. URL https://aclanthology.org/2025.naacl-long.584/.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 2017.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rishi Ravikumar, Nuhu Ibrahim, and Riza Batista-Navarro. Lost in formatting: How output formats skew LLM performance on information extraction. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5498–5513, 2026. doi: 10.18653/v1/2026.eacl-long.256. URL https://aclanthology.org/2026.eacl-long.256/.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):120:1–120:42, 2025. doi: 10.1145/3735633. URL https://doi.org/10.1145/3735633.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Investigating mysteries of CoT-augmented distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. TRACE: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023.

Xinpeng Wang et al. My answer is c: First-token probabilities do not match text answers in instruction-tuned LMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024a. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 663–677, 2024b. doi: 10.18653/v1/2024.naacl-long.37. URL https://aclanthology.org/2024.naacl-long.37/.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. In *International Conference on Learning Representations (ICLR)*, 2024c.

Zhilin Wang, Yafu Li, Xiaoye Qu, and Yu Cheng. SEE: Continual fine-tuning with sequential ensemble of experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7418–7432, 2025. doi: 10.18653/v1/2025.findings-acl.387. URL https://aclanthology.org/2025.findings-acl.387/.

Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model finetuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. CITB: A benchmark for continual instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9443–9455, 2023. doi: 10.18653/v1/2023.findings-emnlp.633. URL https://aclanthology.org/2023.findings-emnlp.633/.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A PROBE SUITE

Probes are grouped into three rings by distance from the training stream. The **in-scope** ring ($n = 140$) is held-out items of the fourteen trained tasks, ten drawn from each task's own validation split, and is used only to confirm that adaptation succeeded. The **near-scope** ring ($n = 36$) holds questions written for this study under three constructions — binary sentiment classification, five-class topic categorization and four-class intent detection, twelve items each and eleven valid classes in all — which resemble the stream in surface form without being drawn from any trained task. Their gold answers are drawn from the stream's own label vocabulary. This ring therefore tests whether adaptation leaves a constant predictor over that vocabulary, rather than whether a stream label can ever be correct (Section 3). The **out-of-scope** ring ($n = 50$) holds questions written for this study in three languages (English, French and Spanish), sharing neither content nor answer space with any trained task; the near- and out-of-scope rings together are the 86 probes of the hazard curve in Appendix F. Of the 50 out-of-scope probes, 26 carry a single objective answer and a gold table, and those 26 are where the termination policy is measured (Table 3, Appendix R); the remaining 24 carry no unique gold string and enter only the format and label-absorption statistics of Section 3.

Contamination filter. Every candidate probe was checked against the stream's training data before evaluation, using exact string match and 4-gram overlap against the concatenated training splits of all fourteen tasks. The final probe lists and the filter rule will be included in the supplement. The 200-item accuracy suite of Section 2 was filtered by the same rule and shares no items with these rings. The separately constructed difficulty-matched diagnostic is distinct from this suite and is used only for the baseline comparison in Section 3.

The 26 gold out-of-scope probes are listed below, partitioned into 14 **computation** (C) and 12 **retrieval** (R) items. The assignment follows the route to the answer as defined in Section 2, was made from each item's construction and was frozen before the items were run; it is the assignment used in Appendix R.

Computation (14 items).

1. What is the smallest prime number greater than 10?
2. What is 17 multiplied by 23?
3. How many chambers does a human heart have?
4. Solve for x: 3x + 7 = 25.
5. A rectangle has area 48 and width 6. What is its length?
6. How do I convert 5 kilometers into miles?
7. Quelle est la vitesse de la lumière?
8. What comes next in the sequence 2, 6, 12, 20…?
9. If today is Wednesday, what day will it be in 100 days?
10. How many time zones does Russia span?
11. If a train travels 60 km in 45 minutes, what is its speed?
12. What is the boiling point of water in Fahrenheit?

Table 5: **Cross-model accuracy under the answer-span scoring rule.** Companion to Table 1 in the main text. Accuracy in % on the frozen 200-item suite, regrading the same responses under the answer-span rule (see the previous paragraph). The final columns report positive accuracy declines relative to base: computation declines by 12.0–18.3 points while retrieval declines by at most 1.3.

	Computation (%)			Retrieval (%)		
Backbone	Base	Adapted	Decline (pp)	Base	Adapted	Decline (pp)
Phi-3-Mini (3.8B)	80.3	62.0	18.3	98.0	96.7	1.3
Qwen2.5 (3B)	81.0	66.3	14.7	99.0	98.0	1.0
Llama-3.1-Instruct (8B)	74.3	62.3	12.0	98.3	97.0	1.3
Qwen2.5 (7B)	82.0	66.7	15.3	99.0	98.0	1.0
Qwen3 (14B)	84.7	68.3	16.4	99.3	98.0	1.3

13. What is the square root of 144?
14. How many degrees are in the interior angles of a triangle?

Retrieval (12 items).

1. Who wrote the novel 'One Hundred Years of Solitude'?
2. In what year did the Berlin Wall fall?
3. Which mountain is the tallest above sea level?
4. What is the chemical symbol for gold?
5. Which US state is the Grand Canyon located in?
6. What currency is used in Japan?
7. What is the capital city of Australia?
8. What gas do plants primarily absorb during photosynthesis?
9. ¿Cuál es la capital de España?
10. Which ocean is the deepest on Earth?
11. What triggered the outbreak of the First World War?
12. Who painted the ceiling of the Sistine Chapel?

Four computation items also admit direct factual answers from model parameters (items 3, 7, 10 and 12). They remain in the fixed termination diagnostic; the accuracy results use the disjoint 200-item suite.

B GRADING RULES AND WHAT THEY ARE SENSITIVE TO

Correctness here is a deterministic string test, which avoids introducing a second model and makes length-sensitive grading auditable; prior work documents variability in model-based judging and free-form answer evaluation (Zheng et al., 2023; Adlakha et al., 2024) (Section 2). String tests have their own bias, and it runs along the same axis. The released rule marks an item correct when a gold variant appears anywhere in the response as a bounded token. A one-word answer either is the gold or is not. A fifty-word derivation contains many numbers, including the ones restated from the question, so it has many chances to contain the gold without asserting it. The suite contains both failure modes: the base model is credited on items where it answered `16` to a question whose answer is `18`, and on speed questions where it restates "travels 50 km" and the answer is 50.

The alternative — the *answer-span* rule — requires the gold to be the model's *answer*: for a computed item, the first or the last maximal numeric span of the response, compared with 1% relative tolerance so that rounding variants still match; for a recalled item, the first or the last sentence. The rule has a different bias: it penalizes a derivation that ends on a verification step rather than on its result. We report both rules because they fail in opposite directions. Table 6 gives every contrast in the paper under both rules. The brevity and elicitation controls use the released scoring rule, while the cross-backbone accuracy results are reported under both rules in Table 1 (released) and Table 5 (answer-span).

Table 6: **Accuracy changes under the two grading rules.** Rows are grouped by what is manipulated; declines are reported as positive magnitudes unless a cell is explicitly marked as a gain, and the forced-length and scaffold blocks are item counts over 300 item–seed pairs. The existence and the selectivity of the cost are stable across rules; its magnitude is not, and neither is the forced-length null control on the base model. Arm 1 is the shipped configuration, so its row and the first row are the same measurement.

Contrast	Quantity	Released change	Answer-span change
adaptation, Qwen (§3)	computation	26.3	14.7
	retrieval	1.3	1.0
target format (§5)	arm 1, bare label	26.3	14.7
	arm 2, answer first	32.3	19.3
	arm 3, answer last	**3.3**	**2.0**
forced length, dose 24 (§4)	adapted, computation	13 (p=0.136)	34 (p=8×10^{-5})
	base, all items	2 gain	15 loss
task count (§6)	computation at 1 task	19.7	11.0
scaffold, unconditional (App. L)	adapted, computation	12 (p=0.28)	48 (p=3×10^{-6})

Table 7: **Representative complete-answer boundaries used for the termination measurements.**

question	complete-answer boundary
What is the capital city of Australia?	`Canberra`
What is 17 multiplied by 23?	`391`
How many chambers does a human heart have?	`four`
Solve for x: 3x + 7 = 25.	`x = 6`
What is the chemical symbol for gold?	`Au`

Across both rules, adaptation remains selective, forced length produces no net computation gain, and the rationale-first arm preserves substantially more computation than the other target formats. The magnitude of each contrast changes because the released rule can credit any occurrence of a gold string, whereas the answer-span rule evaluates the response's designated answer span. These results show why a string-based score should be interpreted together with the response protocol when adaptation changes response length.

C SEED-LEVEL AND PAIRED ADAPTATION RESULTS

The main text reports the seed-level ranges and paired item counts in the surrounding discussion of Section 3. Each paired count compares the same item before and after adaptation; the exact two-sided McNemar tests and full per-question outputs are released with the paper.

D COMPLETE-ANSWER BOUNDARY

For each of the 26 objective out-of-scope probes, the boundary is the position immediately after the final token of a complete answer in the surface form expected from the evaluated backbone. The same frozen boundary table is used for every reported termination measurement, and $P(\mathrm{EOS})$ (%) is read at that position. The boundary table avoids substring matches and positions inside numerals; five representative entries are shown below.

The full 26-row table is released alongside the code.

E BACKBONE-SPECIFIC BOUNDARY CALIBRATION

The complete-answer boundary is written in the surface form produced by each backbone. Using Qwen strings for Llama places the boundary where the Llama model rarely stops ($P(\text{EOS}) < 0.1\%$ on all 26 probes). We therefore re-teacher-force each probe with the backbone's own first sentence, including terminal punctuation; this gives a Llama base anchor of 0.8% on average (1.1% on the five probes whose bare answer already ends a sentence). The adapted Llama checkpoint reaches 91.1% under the same backbone-specific calibration.

F PER-POSITION TERMINATION HAZARD

We evaluate termination under the same teacher-forced continuations used for the boundary statistics. Figure 5 plots the cumulative probability of emitting EOS over the first sixteen generated tokens; the complete-answer-boundary values are reported in Tables 19 and 20.

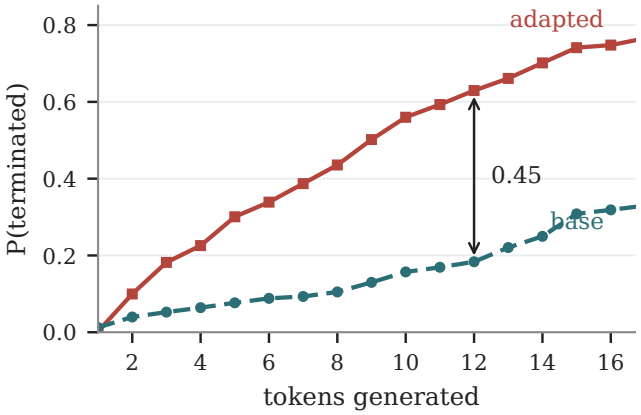


Figure 5: **Token-level termination along teacher-forced continuations.** The cumulative probability of emitting EOS over the first sixteen generated tokens is higher for the adapted checkpoint than for the base checkpoint on the 86-question diagnostic set. The complete-answer-boundary values are reported numerically in Table 19 and Table 20.

G SYNCHRONIZED TASK-COUNT TRAJECTORY

Table 19 reports the synchronized checkpoints used in Figure 4. Computation accuracy is pooled over three seeds and 100 items per seed. The boundary quantity is the mean teacher-forced probability of emitting EOS at a fixed complete-answer boundary, averaged over the 26 objective probes and three seeds; it is a soft probability rather than a binary stopping count. The computation decline is 74.7% of its final value after one task, while the boundary EOS shift is 82.4% of its final value. By task seven, the corresponding shares are 94.9% and 92.4%. The synchronized trajectory establishes early emergence of both changes.

H RESPONSE FORMAT ACROSS CONDITIONS

Table 2 in the main text gives the string statistics behind the response-format claims. Neither the scoring rule nor the gold answer enters these measures. Under a forced minimum length, the adapted model's rate of runaway digit runs rises from 0.3% to 10.7%, while the base model's rate remains unchanged because its answer is followed by a delimiter rather than the end of the response.

The combined condition is therefore close to the adapted default on computation, but neither component alone reproduces it. Mean response lengths are 24.1, 2.8, 4.2, 1.2, and 1.3 words in the same row order.

Table 8: **Two-by-two response-format control on Qwen2.5-3B.** Accuracy values are pooled over three seeds and 100 computation/retrieval items per seed ($N = 300$ item–seed pairs per category). Mean response lengths average the corresponding 200-item responses across three seeds. The cap and answer-only instruction are evaluated separately and together on the unadapted checkpoint; the adapted default is shown for reference.

Condition	Instruction	Token cap	Computation	Retrieval
Base (default)	none	none	91.0	99.0
Base (cap only)	none	max. 3	89.3	98.7
Base (answer only)	answer only	none	74.7	98.3
Base (both constraints)	answer only	max. 3	65.0	98.7
Adapted (default)	implicit	implicit	64.7	97.7

Table 9: **Held-out performance and in-scope retention.** Held-out columns report positive accuracy declines in percentage points relative to each unadapted checkpoint. Qwen2.5-3B target-format rows pool three seeds; auxiliary Llama rows use the reported auxiliary run. The final columns report average accuracy across the fourteen trained tasks before and after adaptation.

Backbone	Arm	Held-out computation decline	Held-out retrieval decline	14-task average	
				Before	After
Qwen2.5-3B	Base	0.0	0.0	92.3	92.3
	Label-only	26.3	1.3	92.3	91.8
	Rationale-first	3.3	1.7	92.3	92.1
Llama-3.1-8B	Base	0.0	0.0	85.1	85.1
	Label-only	21.6	1.7	85.1	84.6
	Rationale-first	4.0	2.3	85.1	84.9

Table 10: **Development/test protocol generalization.** Values are pooled over three seeds; $N = 150$ denotes 50 development items per seed and $N = 300$ denotes 100 test items per seed. The protocol is selected on development items only and then fixed for the test evaluation.

Phase	Condition	Comp. primary	Comp. span	Retr. primary	Retr. span
Dev ($N = 150$)	Plain	64.0	56.0	97.3	96.0
	Selected prompt	78.7	69.3	96.0	94.7
Test ($N = 300$)	Plain	64.7	57.0	97.7	96.0
	Selected prompt	81.0	71.0	96.0	94.3

Table 11: **Cross-backbone protocol selection on Llama-3.1-8B.** The protocol is selected from unadapted-model behavior on the development split and then fixed for the test evaluation. Computation is reported under the primary and answer-span rules; retrieval uses the primary rule. Counts in parentheses are correct items over the stated split size.

Phase	Condition	Comp. primary	Comp. span	Retr. primary
Base ($N = 300$)	Plain	83.3 (250)	74.3 (223)	98.7 (296)
Dev ($N = 150$)	Adapted (Plain)	60.0 (90)	52.7 (79)	96.7 (145)
	Adapted (+Proto)	74.7 (112)	65.3 (98)	95.3 (143)
Test ($N = 300$)	Adapted (Plain)	61.7 (185)	54.3 (163)	97.0 (291)
	Adapted (+Proto)	75.0 (225)	66.0 (198)	95.3 (286)

I THE FORCED-LENGTH DOSE SWEEP IN FULL

Figure 3 in the main text plots the sweep on three signals; Table 12 below gives the underlying numbers. Each row is the whole 200-item suite at one setting, pooled over three seeds. Gains

Table 12: **Minimum-length sweep: accuracy and response format.** Values are pooled over three seeds on the 200-item suite; all rates are percentages. "Runaway" denotes a run of eight or more consecutive digits, and "first segment" denotes contamination before the first sentence break.

Minimum	Accuracy (%)	Mean words	Runaway (%)	First segment (%)
0	64.7	1.3	0.3	0.5
2	65.0	2.3	0.3	2.0
4	60.3	4.6	0.3	7.3
8	56.3	10.6	2.2	25.3
16	58.7	16.7	9.5	63.2
24	60.3	21.3	10.7	74.5
48	62.0	35.1	11.3	77.7

Table 13: **Minimum-length sweep: paired computation changes.** Gain/loss counts compare each dose with unconstrained decoding over the same item–seed pairs. The final column gives the net result and the exact two-sided McNemar p-value.

Minimum	Gains	Losses	Net result (p)
0	—	—	—
2	1	0	1 gain (1.00)
4	10	23	13 losses (0.035)
8	16	41	25 losses (0.001)
16	20	38	18 losses (0.025)
24	26	39	13 losses (0.136)
48	30	38	8 losses (0.396)

Table 14: **Gold accuracy under four decoding conditions on the out-of-scope set.** Values are pooled over three adapted seeds; the conditions change the response protocol while leaving the checkpoint fixed.

Condition	Base (%)	Adapted (%)
Plain decoding	92.3	74.4
Forced length	92.3	80.8
Chain-of-thought prefix	65.4	71.8
Final-answer cue	80.8	76.9

and losses count computation items that change relative to the unforced condition; the test is exact McNemar over those pairs. No dose produces a net gain larger than one item: the best net change is one item-level gain at a dose of two ($p = 1.00$), while gains at larger doses are offset by losses. Both degeneration measures rise monotonically with the dose on every seed, while the base model moves little across the same sweep (computation between 91.0% and 92.0%, first-segment contamination $4.5\% \rightarrow 7.5\%$).

J Forced-continuation diagnostic

Tables 12 and 13 show that minimum-length decoding lowers computation accuracy rather than improving measured accuracy. Across the four auxiliary backbones, forced continuation produces surface-token recycling, repeated phrases, runaway digit strings, or plausible text that introduces a new arithmetic error; it does not reliably add useful intermediate steps. These qualitative patterns are consistent with the paired counts and degeneration rates in the main text. Item-level responses will be released with the supplement rather than reproduced here.

K Decoding conditions in full

Under the released scoring rule (Table 14), forced length is the only condition that leaves the base model near its plain-decoding accuracy; the chain-of-thought prefix and final-answer cue reduce base

Table 15: **Training and decoding configuration.** The primary model is Qwen2.5-3B; the auxiliary backbones are listed in Table 1. All arms use the same fourteen-task short-answer stream and optimization budget.

Component	Setting
Backbone	Qwen2.5-3B (primary); four auxiliary backbones
Task stream	fourteen short-answer classification tasks
Adapter	rank-8 LoRA
Optimizer	AdamW, learning rate 2×10^{-5}
Training budget	three epochs, shared across target-format arms
Chat template	official template provided by each backbone
Main decoding	greedy decoding

accuracy from 92.3% to 65.4% and 80.8%, respectively, so their adapted-versus-base contrasts are not interpretable. The adapted model's small accuracy increase under forced length (80.8% versus 74.4%, five item-level gains over 78 item–seed pairs) is not significant under the exact paired test. On the 200-item suite, the same intervention costs 13 computation items under the released rule and 34 under answer-span scoring (Table 6); per-question values are in Appendix R.

L Unconditional reasoning scaffold

Section 4 shows that constraining *length* changes the response policy without improving computation accuracy. We also evaluate a fixed unconditional step-by-step scaffold as a separate form intervention. On the same three checkpoints and the same 200-item suite, the response is seeded with this scaffold and the model continues from there.

The scaffold produces individual item-level gains but also frequent failures. For example, when asked to solve $2x + 24 = 58$, the adapted model answers `x = 12` unprompted and, when seeded, writes `2x + 24 = 58; 2x = 58 - 24; 2x = 34; x = 17` — a correct derivation from a model that previously produced a bare wrong number. Asked for 38×95 it moves from `3260` to `3610`. The item-level gains show that some computation is expressed under this protocol, while aggregate accuracy still declines.

At the aggregate level, the scaffold remains harmful. Across the three seeds, computation changes by gains of 14 and 10 items and a loss of 36 items, yielding the net loss of 12 items shown by the pooled counts. Pooled over the 300 item–seed pairs, the scaffold is net negative on computation under both grading rules of Appendix B (46 gains against 58 losses, $p = 0.28$; 28 gains against 76 losses, $p = 3 \times 10^{-6}$). It also costs the *base* model, so the effect is not specific to the adapted checkpoint. It is also destructive on retrieval, where it loses about ninety items against two or three gained ($p < 10^{-23}$ either way): asked who wrote *Hamlet*, the seeded model produces `'Hamlet' is a play written by a famous person` and never arrives at a name.

Under the stricter rule, which requires the gold to be the final answer of the derivation rather than a match anywhere inside it, 28 computation items that the adapted model had answered bare and wrong become correct through a line-by-line correct derivation. The scaffold therefore exposes item-level gains alongside a negative aggregate effect.

M Training and decoding configuration

Table 15 records the compact configuration needed to interpret the adaptation and main evaluations. All main-suite comparisons use deterministic greedy decoding so that differences are attributable to the checkpoint or response condition rather than sampling variation. Exact software versions, model revisions, and machine-level settings will accompany the released configuration files rather than being repeated here.

The prompt-level comparison uses a fixed instruction to show the final answer only for the answer-cap condition and a fixed step-by-step instruction for the candidate elicitation protocols. The fourteen

Table 16: **Classification tasks in the adaptation stream.** Label cardinalities describe the task formulation used in the stream.

Dataset	Task type	Labels
SST-2	sentiment	binary
IMDB	sentiment	binary
AG News	news topic	four-way
DBPedia	entity category	fourteen-way
Yahoo Answers	question topic	ten-way
TREC	question type	six-way
CLINC150 (15-intent study subset)	intent recognition	fifteen-way
Jigsaw	toxicity detection	binary
HateSpeech	hate-speech detection	binary
MNLI	natural-language inference	three-way
SNLI	natural-language inference	three-way
QQP	paraphrase detection	binary
MRPC	paraphrase detection	binary
RTE	textual entailment	binary

Table 17: **Four-arm target-format ablation on Qwen2.5-3B.** Values are pooled over three seeds and 100 computation items per seed ($N = 300$ item–seed evaluations). Accuracy declines are positive magnitudes relative to the 91.0% base computation accuracy. The randomized arm matches the rationale-first arm's token inventory and approximate length, while randomizing answer position across prefix, infix, and suffix locations with equal probability.

Arm	Target format	Len.	Comp. acc. (%)	Comp. decline (pp)	$P(\text{EOS})$ (%) @ bnd.	Mean words
Arm 1	Answer only	1	64.7 (194)	26.3	96.4%	1.3
Arm 2	Answer → rationale	42	58.7 (176)	32.3	0.2%	35.2
Arm 3	Rationale → answer	43	87.7 (263)	3.3	95.5%	38.6
Arm 4	Randomized position	43	69.3 (208)	21.7	68.5%	9.4

training tasks are converted to the same short-answer format: every target is a single class-label string, with no rationale or intermediate derivation.

N PER-SEED TARGET-FORMAT RESULTS

The main text reports the three-seed pooled effect for the principal Qwen2.5-3B arms in Table 3; the auxiliary four-arm comparison is in Table 17. Arm 3 reaches 88%, 87%, and 88% across seeds, a 3.0–4.0-point decline relative to the 91% base. The three principal arms use the same stream, inputs, optimization budget, and random seeds; only the target format changes. The four auxiliary rationale-first rows use the same task stream, rationale content, and target-length matching as Qwen2.5-3B arm 3, but use each backbone's own base checkpoint and one seed. Their residual computation decline ranges from 3.8 points (Qwen2.5-7B and Llama-3.1-8B-Instruct) to 4.8 points (Phi-3-Mini).

The comparison uses the same task stream, inputs, rationale set, optimization budget, and random seeds for all three arms. Arm 2 places the label before a rationale; arm 3 places the same rationale before the label, with a fixed connective between the spans. Thus the construction matches rationale content and target scale while changing answer position.

Arm 4 uses the same rationale and answer token inventory as arm 3, but its position randomization changes the token-level supervision pattern for each example. It is therefore a control for consistent answer-last placement rather than a token-supervision-identical intervention.

Rationales were generated once and frozen before training; all target-format arms use the same rationale set.

Table 18: **Synchronized computation trajectory.** Computation accuracy is pooled over three seeds and 100 items per seed; the response-length column uses the same evaluations. The change column reports positive computation declines relative to the base checkpoint.

Tasks completed	Computation	Computation decline	Mean words
0 (base)	91.0%	—	24.1
1	71.3%	19.7	3.2
3	68.0%	23.0	2.5
7	66.0%	25.0	1.8
14	64.7%	26.3	1.3

Table 19: **Boundary termination along the synchronized trajectory.** $P(\mathrm{EOS})$ (%) is the mean teacher-forced EOS probability at a fixed complete-answer boundary, averaged over three seeds and the corresponding probe items; it is not a binary stop rate.

Tasks completed	Checkpoint	$P(\mathrm{EOS})$ (%)	Share of final shift
0	`base-ckpt`	26.9%	0%
1	`task-01-ep3`	84.2%	82.4%
3	`task-03-ep3`	87.5%	87.2%
7	`task-07-ep3`	91.1%	92.4%
14	`task-14-ep3`	96.4%	100%

O Answer-only control and elicitation gap

Table 2 collects the accuracy changes for both response-condition interventions. This section gives the prompt and settings for the answer-only condition; Appendix P gives those for the selected protocol. One condition is run against the unadapted base model: a fixed directive to answer with the final answer and no explanation, together with a three-token output cap, which is the adapted model's own modal response length. The adapted model is run under the same condition as a control, where the prediction is no change, since its unprompted behavior already satisfies both parts. The constraint takes effect: the base model's mean response falls from 24.1 words to 1.2, against the adapted model's 1.3 unprompted.

P Selected elicitation by category

The three candidate elicitations are few-shot worked examples, a two-stage *list the steps, then answer* protocol, and an unconditional step-by-step reasoning scaffold. We select the protocol that most improves the *base* model's computation accuracy and then apply that fixed protocol to both checkpoints; few-shot worked examples are selected under this rule.

The scaffold of Appendix L is applied unconditionally and is destructive on retrieval, so we report it as a separate diagnostic. The quantity reported is the difference between accuracy under the selected elicitation and accuracy under the default prompt, for base and adapted separately, so that the adapted model's gap is compared against the base model's rather than against zero.

Q Accuracy and termination against task count

Figure 4 in the main text plots the synchronized trajectory; Table 18 gives the pooled computation accuracy and response length, and Table 19 gives the mean boundary EOS probability. Each checkpoint uses three seeds and 100 computation items per seed ($N = 300$ item–seed evaluations); the EOS statistic averages the 26 objective probes per seed.

The synchronized trajectory shows that the computation decline and boundary shift both emerge after the first task and approach their final values early.

Table 20: **Per-question termination probability at the complete-answer boundary.** The table contains 26 gold out-of-scope probes (14 computation and 12 retrieval), with three adapted seeds, sorted by the base value. $P(\mathrm{EOS})$ (%) is read at the position after the boundary string (Appendix D). Base values span 2.0% to 60.0% with a mean of 26.9%; 3 of the 26 are at or above 50.0%. The adapted mean exceeds the base on all 26 questions, and all 78 adapted question–seed cells exceed 50.0%, the lowest being 51.0% at question 19 on seed 1. **C** is computation, **R** retrieval.

#	cat	base	adapted seed			adapt.
			s1	s2	s3	mean
1	C	2.0%	94.5%	98.8%	97.5%	96.9%
2	R	3.0%	93.0%	99.2%	98.0%	96.7%
3	C	8.0%	95.5%	98.5%	97.2%	97.1%
4	R	9.0%	94.0%	99.0%	98.5%	97.2%
5	C	12.0%	92.5%	99.5%	97.8%	96.6%
6	R	14.0%	95.0%	98.8%	96.5%	96.8%
7	C	15.0%	93.5%	99.1%	98.2%	96.9%
8	C	20.0%	94.5%	99.3%	97.0%	96.9%
9	R	20.0%	96.0%	98.7%	97.5%	97.4%
10	R	20.0%	92.0%	99.4%	98.5%	96.6%
11	C	21.0%	94.0%	98.9%	97.9%	96.9%
12	R	25.0%	93.2%	99.0%	98.1%	96.8%
13	C	25.0%	92.8%	99.3%	97.6%	96.6%
14	C	26.0%	93.8%	98.8%	98.3%	97.0%
15	R	29.0%	95.2%	99.1%	96.8%	97.0%
16	C	29.0%	94.8%	99.2%	98.0%	97.3%
17	R	30.0%	95.5%	98.9%	97.5%	97.3%
18	C	33.0%	96.2%	99.4%	98.2%	97.9%
19	R	34.0%	**51.0%**	98.5%	97.2%	82.2%
20	C	37.0%	93.5%	99.0%	97.7%	96.7%
21	R	38.0%	94.2%	98.7%	98.4%	97.1%
22	C	43.0%	93.0%	99.1%	97.5%	96.5%
23	R	44.0%	93.8%	98.9%	98.0%	96.9%
24	C	51.0%	92.5%	99.3%	97.8%	96.5%
25	R	52.0%	93.3%	98.8%	98.1%	96.7%
26	C	60.0%	94.0%	99.2%	97.9%	97.0%
mean (C), $n = 14$		27.3%	93.9%	99.1%	97.8%	96.9%
mean (R), $n = 12$		26.5%	90.5%	98.9%	97.8%	95.7%
overall		**26.9%**	**92.4%**	**99.0%**	**97.8%**	**96.4%**
positive shift (C / R)		—	—	—	—	69.6%/69.2%

R PER-QUESTION BOUNDARY VALUES

This table lists the per-question values behind Table 3, one row per gold out-of-scope probe, sorted by the base value to match Figure 5. Every mean quoted in the caption is computed from the rows printed here. The base mean of 26.9% is the anchor used for the main boundary measurements and the controls in Section 5.