When Unfinished Counts as Wrong: Generation Budgets Define the Target in Reasoning Monitor Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Reasoning monitors estimate whether an ongoing reasoning trajectory will eventually yield a correct answer, enabling early stopping, verification, and runtime oversight. Many evaluations obtain the target label by matching answers from generations produced under a fixed token budget. Under a larger-budget prediction objective, however, this procedure can assign the same negative label to completed errors and trajectories that simply ran out of budget before producing their intended final answer. Our key observation is that the generation budget is therefore part of the prediction target, not merely an implementation detail. We audit this dependence across seven open models and 19 model–dataset configurations, comparing 256-token labels with paired 2048-token operational reference labels for 2,886 trajectories while holding monitor scores fixed. Across configurations, heuristic truncation reaches up to 92.5% and short-to-reference label agreement falls as low as 28.5%; point-estimate model rankings can also change. On DeepSeek-R1-7B/GSM8K, a last-step probe scores 0.892 AUC against short-budget labels but 0.568 against the reference labels. Budget sweeps and completion controls are consistent with completion-related confounding, whereas reference-label retraining yields limited, configuration-dependent recovery. These results expose a target-validity failure mode in reasoning-monitor evaluation and motivate paired label auditing before monitor performance is interpreted as longer-horizon correctness prediction.

1 Introduction

Reasoning monitors aim to estimate whether an ongoing reasoning trajectory will eventually yield a correct outcome before generation is complete. Such estimates can support early stopping, verification, compute allocation, and runtime oversight, where a system must decide whether to continue, trust, or intervene on a trajectory (Chan et al., 2025; Yuan et al., 2026; Chrabąszcz et al., 2026). Yet monitor performance is meaningful only relative to the target being predicted. In evaluations that derive correctness by matching a model-generated answer to a reference answer, the generation procedure is therefore part of the measurement design: it helps define the label against which the monitor is judged.

This dependence becomes important when reasoning is bounded by a finite output budget, since reasoning traces can extend well beyond short generation caps (Chen et al., 2026; Sui et al., 2025). If generation stops before the intended final answer is produced, the extractor may return no answer or may fall back to an intermediate value. A binary label can then assign the same negative outcome to two qualitatively different cases: a completed trajectory that reaches an incorrect answer, and an unfinished trajectory that might succeed if allowed to continue. Whether this is a labeling error depends on the intended target. If deployment defines success as answering correctly within the same fixed budget, an unfinished output is legitimately a failure; if the monitor is intended to predict correctness under a larger generation allowance, the same short-budget label can misrepresent the outcome of interest.

Recent work assesses reasoning quality with process reward models, hidden-state probes, and temporal predictors of eventual outcomes (Lightman et al., 2024; Wang et al., 2024; Zhang et al.,

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.

2025a; Sun et al., 2025; David, 2025; Yuan et al., 2026; Khalifa et al., 2026), but these studies do not all construct targets the same way: some use controlled or externally supervised labels, whereas others determine correctness from self-generated answers under finite generation limits. Work on test-time reasoning further shows that reasoning length, stopping behavior, and token allocation materially affect outcomes (Muennighoff et al., 2025; Ghosal et al., 2025; Wu et al., 2025; Yi et al., 2025). For evaluations that combine bounded self-generation with answer matching, this creates a measurement question distinct from improving the monitor itself: how stable are the labels, and the conclusions drawn from them, to the budget used to construct the outcome?

Our key insight is that the generation budget is part of the prediction target, not merely an implementation detail. A label produced at budget B answers whether the observed generation is correct under that procedure; it need not answer whether the same model and prompt would be correct with a larger allowance. This changes how high monitor AUC should be interpreted: if many short-budget negatives are unfinished but recoverable, completion, answer readability, or termination-related features can correlate strongly with the short-budget target, so a monitor may discriminate short-budget labels well without equivalently predicting longer-horizon correctness. The relevant question is not simply whether a monitor predicts its labels, but whether those labels represent the capability the evaluation claims to measure.

We study this issue with a paired reference-label audit. For each sampled prompt, we compare the original label obtained with a 256-token generation budget with an operational reference obtained by regenerating the same model and prompt with a 2048-token budget while preserving the answer-extraction and matching procedure. We then evaluate the same cached monitor scores against both label sets. Holding scores fixed isolates sensitivity to label construction from changes in monitor training or representation. The audit additionally records heuristic completion indicators, native termination information, and paired-prefix evidence, and uses larger-budget checks to determine when the 2048-token reference itself remains unstable. We apply this protocol to seven open models across MATH, GSM8K, and HotpotQA, covering 19 model–dataset configurations and 2,886 paired trajectories; ALFWorld contributes an additional truncation analysis.

The resulting label shifts are large in several configurations. At 256 tokens, heuristic truncation reaches 92.5%, while agreement with the 2048-token operational reference falls to 28.5%. Across the 2,886 paired examples, 868 labels change from negative to positive and 60 change in the reverse direction, and the point-estimate model ranking on GSM8K changes with the labeling budget. Monitor conclusions shift as well even though the scores are unchanged. On DeepSeek-R1-7B/GSM8K, the last-step probe has an AUC of 0.892 against the 256-token labels but 0.568 against the reference labels; TDE changes from 0.830 to 0.524 and Pnorm from 0.878 to 0.496. Across the core analysis, 17 of 33 tested AUC decreases remain significant after Bonferroni correction.

Several controls clarify what these changes do and do not imply. Completion-only baselines, paired-label quadrant analyses, and answer-presence controls are consistent with completion-related confounding under severe truncation, but they do not identify a single representation-level mechanism, and retraining on reference labels yields only limited, configuration-dependent recovery. Moreover, the 2048-token label is not unrestricted ground truth: some configurations remain incomplete at 4096 or 8192 tokens, so we retain an unresolved status rather than treating a larger budget as automatically definitive. Our claim is therefore about evaluation validity under bounded label construction, not a general failure of reasoning monitoring.

Our contributions are threefold. **(i) Target formulation:** we formalize a generation-budget-induced target mismatch, where bounded answer-matching labels conflate unfinished trajectories with completed errors when the intended target is correctness under a larger allowance, so the budget is part of the target definition. **(ii) Paired empirical audit:** we introduce a fixed-score reference-label audit and show that changing the label-generation budget alters measured task accuracy, point-estimate model rankings, and monitor AUC across seven models, 19 configurations, and 2,886 trajectories, with a six-level budget sweep characterizing how truncation and agreement evolve. **(iii) Diagnosis and reporting protocol:** we combine completion baselines, label decompositions, retraining controls, and larger-budget stability checks to diagnose when the effect appears and preserve uncertainty when the reference remains incomplete, and we will release RM-RES with paired labels, monitor scores, completion indicators, and reference-status metadata.

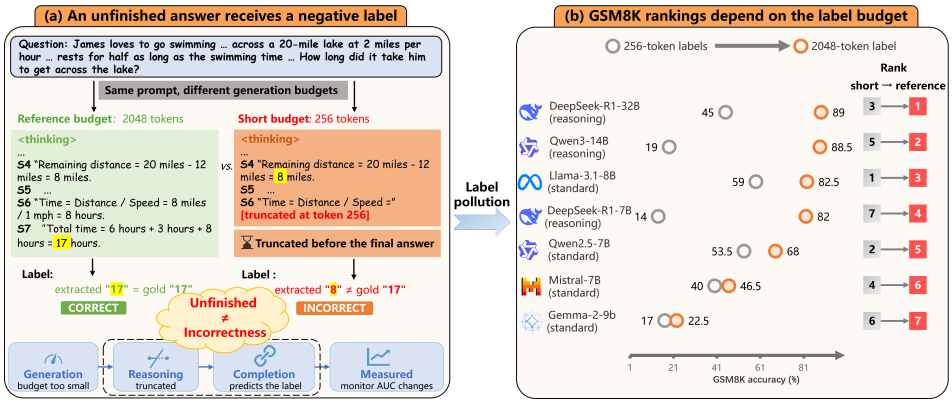

Figure 1: The labeling budget changes what an evaluation counts as success. (a) A paired DeepSeek-R1-7B/GSM8K example. The 256-token output stops before the intended final answer, and the extractor returns the intermediate value 8. The 2048-token output reaches the gold answer, 17. The short-budget negative therefore does not represent the larger-budget outcome. (b) GSM8K accuracy under the two budgets for seven models. Each pair represents the same model and prompts. Ranks show point estimates, not established differences between neighboring models.

2 Related Work

Reasoning monitors and their targets. Process reward models assess intermediate steps using human annotations or automatically generated supervision (Lightman et al., 2024; Wang et al., 2024; Khalifa et al., 2026). Hidden-state probes estimate answer or trajectory correctness, while temporal predictors anticipate later outcomes from partial trajectories (Zhang et al., 2025a; Sun et al., 2025; David, 2025; Yuan et al., 2026; Chrabąszcz et al., 2026). Related approaches predict misalignment and agent-episode failure (Chan et al., 2025; Ruan et al., 2026). These tasks differ in both the information available to the monitor and the outcome used as its target. Our study concerns trajectory-level labels extracted from bounded, self-generated answers.

Label construction. Controlled arithmetic targets, human step annotations, and independently determined episode outcomes separate supervision from the monitor's observation budget. In contrast, correctness extracted from a capped generation depends on the procedure that produces it. For example, the protocols examined by Yuan et al. (2026) and David (2025) use finite limits when constructing answer-based outcomes. Appendix A.2 compares fourteen protocols by label source and cap handling, identifying where an audit is relevant without establishing that a published result is affected.

Generation budgets and reasoning outcomes. Prior work studies budget forcing, token-aware generation, overthinking, and adaptive stopping (Muennighoff et al., 2025; Han et al., 2025; Ghosal et al., 2025; Wu et al., 2025; Fu et al., 2025; Yi et al., 2025), and examines adaptive length rewards, competition between reasoning and answer generation, and the separation of completion from conditional correctness (Liu et al., 2026; Nie et al., 2026; Kaiser et al., 2026). We examine how these effects enter the labels used to evaluate monitors; deliberately eliciting a final answer after limiting reasoning is a different protocol from stopping generation mid-reasoning.

Measurement validity and shortcut prediction. Label errors can alter benchmark conclusions, and predictive models can exploit incidental features of an evaluation target (Northcutt et al., 2021; Geirhos et al., 2020; Hewitt & Liang, 2019; Belinkov, 2022). Evaluation and reward-model studies examine protocol and length effects (Chang et al., 2024; Zheng et al., 2023; Gao et al., 2023; Dubois et al., 2024), while process-supervision research addresses noisy step labels (Duan et al., 2025; Xie et al., 2026; Zhang et al., 2025b). Our control changes the budget used to construct outcomes while holding monitor scores fixed, which makes sensitivity to the target directly observable.

3 GENERATION BUDGETS AND THE PREDICTION TARGET

3.1 DEFINING THE OUTCOME

Let x be a prompt, $g(x)$ its gold answer, and $T_B(x)$ the output produced by a fixed model with generation budget B. With answer extractor A and matching rule M, the automatic label is

$$y_B(x) = \mathbb{1}[M(A(T_B(x)), g(x)) = 1]\,. \tag{1}$$

The prompt, model, decoding procedure, extractor, and matcher are part of this definition, so changing the budget changes the output to which extraction is applied. A missing intended answer need not produce an empty extraction: a fallback may return an intermediate value.

The observation horizon and outcome horizon are separate choices. A monitor may score a prefix observed at 256 tokens while predicting an outcome obtained with 2048 tokens. Conversely, if deployment requires a correct answer within 256 tokens, y_{256} is the appropriate target, including failures caused by running out of budget. Target mismatch arises when one outcome is used to evaluate a claim about the other.

We compare $y_s = y_{256}$ with the operational reference $y^* = y_{2048}$, obtained by regenerating the same model and prompt. This is a bounded reference, not unrestricted ground truth. Completion indicators and reference-status metadata are retained separately from extracted correctness.

A paired regeneration is not automatically a continuation of the cached prefix, since greedy runs can diverge under numerical or implementation differences. The comparison is therefore an audit of same-model, same-prompt label construction; Appendix B.3 repeats it on the examples where the cached output is a strict token prefix of its regeneration, which is direct evidence of a shared path.

3.2 WHY UNFINISHED OUTPUTS CHANGE THE NEGATIVE CLASS

Consider the idealized event E that the short cap is reached before the intended final answer, with probability τ. Suppose such an output is never labeled positive and completed outputs retain their labels under the larger budget. Let $p^*_E = P(y^*{=}1 \mid E)$ and $p^* = P(y^*{=}1)$. Then

$$P(y_s{=}y^*) = 1 - \tau p^*_E, \qquad P(y_s{=}0 \mid y^*{=}1) = \frac{\tau p^*_E}{p^*}, \quad p^* > 0. \tag{2}$$

The product τp^*_E is the fraction unfinished at the short cap but positive under the reference. Thus, frequent truncation creates large disagreement when many truncated outputs succeed under the larger allowance. Agreement can remain high when both budgets produce mostly negatives.

Actual extraction, completion indicators, and paired generation can violate these assumptions. Appendix C.1 gives the general relations, including changes outside E. Empirical conclusions use the observed paired labels; quantities inferred from agreement itself are not treated as independent validation of Equation (2).

3.3 THE PAIRED REFERENCE-LABEL AUDIT

For each held-out prompt, we retain the short output and cached monitor score, generate a larger-budget reference, and apply the same extractor and matcher. We report positive-label rates, both directions of label change, agreement, and the reference false-negative rate $P(y_s{=}0 \mid y^*{=}1)$. This rate differs from the proportion of short-budget negatives that become reference-positive.

For a fixed score vector s, we measure

$$\Delta_{\mathrm{AUC}} = \mathrm{AUC}(s, y_s) - \mathrm{AUC}(s, y^*). \tag{3}$$

Both AUCs use identical examples and predictions. Their difference isolates sensitivity to the evaluation labels from changes in monitor training or representation. A positive difference means that the scores discriminate the short-budget target better; it does not identify the feature responsible or the best achievable reference performance.

Reference subsets contain 96–200 examples per configuration. The basic audit requires one larger-budget regeneration per example, without human annotation or retraining. Separate 4096-token checks

assess reference stability, followed by 8192-token escalation for two low-completion configurations; unresolved configurations retain their operational labels and status rather than being assigned a definitive outcome. We use 70% agreement as a descriptive warning threshold, not a calibrated decision rule.

3.4 Generation and completion indicators

We distinguish heuristic completion, native termination, exact-prefix continuation, and final-answer-marker presence. The heuristic examines unfinished syntax and recognized answer markers in the final reasoning step. Native termination records whether the engine stops without reaching its cap. Exact-prefix continuation requires a capped short output to be a strict token prefix of its paired longer output. Marker presence records the explicit `Final Answer` marker. Appendix A.1 specifies the rule and the recognized markers.

These indicators are not interchangeable. A cap may occur after an answer, a complete-looking sentence may precede an unfinished solution, and a valid answer may omit the marker. To distinguish cap events with and without a detected answer, we also report

$$\tau_E = \tau_{\text{cap}} \left[1 - P(\text{answer present} \mid E_{\text{cap}})\right], \tag{4}$$

where E_{cap} is the exact-prefix cap event. This is an operational indicator, not direct observation of semantic truncation. We report these indicators separately because they capture different aspects of truncation.

4 Experimental Setup

Research questions. RQ1 asks how the labeling budget changes trajectory labels and model-level conclusions. RQ2 asks how those changes affect fixed-score monitor evaluation, and RQ3 whether completion and answer availability explain the sensitivity.

Models and tasks. We evaluate DeepSeek-R1-Distill-Qwen-7B and -32B (DeepSeek-AI, 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)[1], Gemma-2-9B-IT (Gemma Team et al., 2024), and Qwen3-14B (Yang et al., 2025). The paired audit covers MATH, GSM8K, and HotpotQA for the five smaller models, and MATH and GSM8K for the two additional models (Hendrycks et al., 2021; Cobbe et al., 2021; Yang et al., 2018): 19 configurations and 2,886 trajectories. ALFWorld contributes truncation statistics without paired interactive regeneration (Shridhar et al., 2021). A separate AIME analysis examines transfer to harder mathematics.

Generation and labels. Prompts request numbered reasoning steps and a final answer. Original trajectories use greedy decoding with a 256-token cap, which we treat as a short-budget stress test rather than a reproduction of any audited protocol; Section 5.1 sweeps 128–4096 tokens, including the 512-token setting used by the audited monitoring protocols, and establishes how far the resulting sensitivity extends. Primary references use 2048 tokens. Dataset-specific extraction obtains numerical answers for GSM8K, boxed expressions for MATH, and normalized entities for HotpotQA. Prompts and extraction procedures remain fixed within each budget comparison. Original monitor-training sets contain 2,000–6,000 trajectories per configuration. Reference subsets come from held-out data without overlap with these training sets. The base instruction and extraction priorities are given in Appendix A.1.

Monitors and controls. We cache step-final states from the last eight transformer layers. TDE scores consecutive top-layer state displacements, Pnorm averages logistic predictions from standardized step-final states, and the last-step probe uses the final observed step's top-layer state. These scores use the available short output, which can end naturally before the cap. We also evaluate ThinkPRM-1.5B using its verification-chain procedure (Khalifa et al., 2026), and untrained heuristic-completion and native-termination baselines. Additional controls examine length, answer presence, training labels, and representations.

[1]Checkpoint `mistralai/Mistral-7B-Instruct-v0.3` (`https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`), which extends the base model with a revised tokenizer and vocabulary.

Table 1: Truncation indicators at a 256-token budget. Entries are heuristic truncation rates (%), with exact-prefix cap rates in parentheses. Heuristic estimates use 200 examples per configuration. For the two additional models, GSM8K cap rates use the 50-example budget-sweep subsets. A dash indicates that the corresponding analysis was not performed.

Model	MATH	GSM8K	ALFWorld	HotpotQA
DS-7B	88 (98)	84 (93)	28 (–)	74 (85)
Llama	84 (100)	28 (99.5)	55 (–)	31 (99)
Qwen2.5	88 (95)	35 (68.5)	7 (–)	12 (26)
Mistral	54 (62)	29 (36)	0 (–)	5 (5)
Gemma	46 (57)	11 (17)	0 (–)	33 (19)
DS-32B†	57 (–)	52.5 (56)	–	–
Qwen3†	92.5 (–)	91.5 (98)	–	–

Estimation and inference. Metrics include ROC AUC, paired-label statistics, Pearson correlations, and ten-bin expected calibration error (ECE). Paired bootstrap analyses use 2,000 resamples and seed 42. The core family contains 45 planned AUC comparisons across 15 configurations and three monitors. Twelve have fewer than ten positives under at least one target and are excluded from testing, leaving 33 tests; the Bonferroni denominator remains 45. The additional-model family contains twelve planned comparisons, nine tested, with denominator twelve. Tests for AUC decreases are one-sided. AUC is undefined when either class is absent.

Training and numerical controls. Fixed five-fold splits keep all observations from a trajectory together and vary training and evaluation labels separately. Original caches and primary references use bf16. Additional checks show that greedy reference labels can change between bf16 and fp16, motivating explicit numerical settings and paired-prefix checks. On 20-example checks, fp16 and bf16 reference labels agree on 95% of GSM8K, 70% of MATH, and 75% of HotpotQA examples; matched bf16 step-final states have cosine similarity of at least 0.9999 to the original cache. We therefore distinguish native cap records from paired-prefix matches and report reference stability separately from numerical reproducibility.

5 Results

5.1 RQ1: How does the labeling budget change the target?

Short outputs frequently appear unfinished. Table 1 reports heuristic truncation and exact-prefix cap rates. Heuristic truncation exceeds 80% on MATH for DS-7B, Llama, and Qwen2.5, and reaches 92.5% on Qwen3/MATH. The distinction between indicators matters: Llama/GSM8K has 28% heuristic truncation but 99.5% exact-prefix caps. Many capped outputs therefore look textually complete; cap hits alone do not identify missing answers.

The same prompts receive different labels. Tables 2–3 compare short-budget labels with the operational reference. Agreement ranges from 28.5% to 99.0%. Across 2,886 examples, 868 labels change from negative to positive and 60 reverse. The reference positive rate is higher in 17 of 19 configurations. Qwen2.5/MATH changes from 2.0% to 38.0%, and Qwen3/MATH from 0.0% to 67.0%.

For DS-7B/GSM8K, 136 of 172 short-budget negatives become positive (79.1%), and the same examples represent 82.9% of the 164 reference positives. Reverse changes also occur, so the reference is not a per-example upper bound on the short-budget label.

Model-level conclusions also change. On GSM8K, DS-7B rises from 14.0% to 82.0%, close to Llama's 82.5%, and DS-32B and Qwen3 rise from 45.0% and 19.0% to 89.0% and 88.5% (Figure 1(b)); small differences among neighboring estimates do not establish resolved rankings.

Label stability varies across budgets. We sweep 128, 256, 512, 1024, 2048, and 4096 tokens on six configurations with 50 paired prompts each. The effect is not specific to the 256-token setting: on DS-7B/GSM8K the short-label to reference-label AUC gap is 0.911→0.457 at 512 tokens, close to its value at 256, and it largely closes by 1024 (0.762→0.730). Truncation generally decreases

Table 2: Reference-label verification (all 15 combinations; bf16 reference; 95% bootstrap CIs on agreement and the reference false-negative rate). p_B/p^*: short-budget/reference positive rates; p_t^*: reference accuracy among outputs judged truncated (Appendix C.1); Completion: reference-side `Final Answer` rate. ‡: reference-side positives < 10—the reference false-negative rate is unreliable there. Extended-model cells in Table 3.

Combination	n	p_B	p^*	Agree.	Ref.-rel. FNR	p_t^*	Compl.
DS-7B/MATH	140	2.9	34.3	67.1 [59.3, 74.3]	93.8 [86.0, 100.0]	0.31	39.3
Qwen2.5/MATH	100	2.0	38.0	64.0 [55.0, 73.0]	94.7 [86.5, 100.0]	0.37	75.0
Llama/MATH	96	6.2	39.6	62.5 [52.1, 71.9]	89.5 [79.1, 97.7]	0.37	55.2
DS-7B/GSM8K	200	14.0	82.0	32.0 [25.5, 38.5]	82.9 [77.2, 88.5]	0.83	95.0
DS-7B/HotpotQA	150	18.0	17.3	76.7 [70.0, 82.7]	65.4 [46.2, 83.3]	0.17	40.7
Mistral/MATH‡	100	1.0	2.0	99.0 [97.0, 100.0]	50.0 [0.0, 100.0]‡	0.00	95.0
Gemma/MATH	100	11.0	16.0	91.0 [85.0, 96.0]	43.8 [18.8, 68.8]	0.14	79.0
Qwen2.5/GSM8K	200	53.5	68.0	77.5 [71.5, 83.5]	27.2 [19.9, 34.8]	0.57	99.5
Gemma/HotpotQA	100	36.0	40.0	92.0 [87.0, 97.0]	15.0 [4.8, 26.5]	0.25	75.0
Llama/HotpotQA	100	45.0	47.0	90.0 [84.0, 95.0]	12.8 [4.2, 23.3]	0.32	74.0
Mistral/GSM8K	200	40.0	46.5	87.5 [83.0, 92.0]	20.4 [12.4, 28.4]	0.29	99.0
Llama/GSM8K	200	59.0	82.5	71.5 [65.0, 78.0]	31.5 [24.3, 38.7]	0.71	96.0
Qwen2.5/HotpotQA	100	43.0	44.0	95.0 [90.0, 99.0]	6.8 [0.0, 15.1]	0.31	97.0
Gemma/GSM8K	200	17.0	22.5	91.5 [87.5, 95.0]	31.1 [17.6, 45.5]	0.47	99.5
Mistral/HotpotQA	100	39.0	35.0	96.0 [92.0, 99.0]	0.0 [0.0, 0.0]	0.00	96.0

Table 3: Paired label audit for DeepSeek-R1-32B and Qwen3-14B. Definitions and confidence intervals follow Table 2. Each configuration contains 200 examples. The 2048-token labels are the primary reference; their stability under larger budgets is reported separately.

Combination	n	p_B	p^*	Agree.	Ref.-rel. FNR	p_t^*	Compl.
Qwen3/GSM8K	200	19.0	88.5	28.5 [22.5, 35.0]	79.7 [73.3, 85.5]	0.90	94.0
Qwen3/MATH	200	0.0	67.0	33.0 [26.5, 39.5]	100.0 [100.0, 100.0]	0.69	83.0
DS-32B/MATH	200	29.5	75.0	54.5 [48.0, 61.5]	60.7 [52.8, 68.0]	0.76	94.0
DS-32B/GSM8K	200	45.0	89.0	55.0 [48.0, 62.0]	50.0 [42.5, 57.1]	0.86	98.5

and reference agreement generally increases as more tokens are allowed (Figure 2, Appendix B.1). Comparisons beyond 2048 test whether the primary reference remains stable, and AUC comparisons that retain the 256-token score are distinguished from those that score an extended output in Appendix B.4.

The larger-budget target still needs validation. In 50-example stability checks, agreement between 2048- and 4096-token labels is 90–100% on tested GSM8K configurations and 96–100% for the additional models, but only 76% on DS-7B/MATH and 62% on DS-7B/HotpotQA; both remain flagged after escalation to 8192 tokens (Appendix B.2). Reference status is reported per configuration throughout; the three headline quantities in Section 1 come from Qwen3/MATH, Qwen3/GSM8K, and DS-7B/GSM8K, whose 2048-to-4096 agreement is 96%, 98%, and 96% respectively.

An adaptive rule selects the smallest tested budget with at least 95% heuristic completion. Where a qualifying budget exists, fixed-prefix AUC is close to its value under the 2048-token reference (Appendix B.4). On AIME, each model has 133 examples, none short-budget positive and eight reference-positive; agreement is 94% despite a 100% cap rate, illustrating that agreement must be read alongside class prevalence (Appendix B.5).

5.2 RQ2: How does changing the target affect monitor performance?

Identical predictions support different performance estimates. Table 4 changes labels without changing monitor scores. On DS-7B/GSM8K, last-step AUC changes from 0.892 to 0.568, TDE from 0.830 to 0.524, and Pnorm from 0.878 to 0.496. High short-budget discrimination therefore does not transfer to the larger-budget target. Across the core family, 21 of 33 tested decreases are significant before correction and 17 remain significant after Bonferroni correction. Several MATH gaps are large descriptively but have too few short-budget positives for reliable testing.

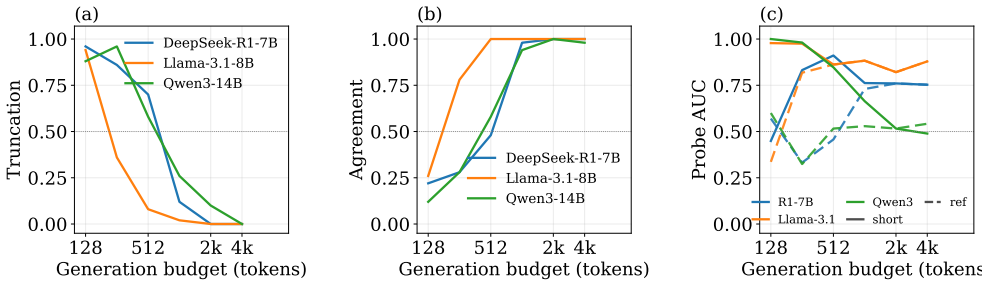

Figure 2: Generation budgets change truncation, label agreement, and measured performance. Three GSM8K configurations use 50 paired prompts each, with the same prompts across budgets. (a) Heuristic truncation rate. (b) Agreement with the operational 2048-token reference. Colors identify models; agreement alone does not establish reference completeness. (c) Last-step probe AUC, $\mathrm{AUC}(s_B, y_B)$ against $\mathrm{AUC}(s_B, y_{2048})$, where s_B is refit at each budget; solid and dashed lines are the two curves, and the two curves coincide at 2048 tokens by construction, since that budget reuses the reference labels; beyond 2048 tokens they can separate again when a budget-specific label differs from the 2048-token operational reference. Figure 3 includes all six configurations.

Table 4: Monitor AUC under short-budget and reference labels. Each entry is $\mathrm{AUC}(y_B) \rightarrow \mathrm{AUC}(y^*)$, evaluated with the same scores on the reference subsets in Table 2. Monitors use the original training sets. Completion-only uses the text-based heuristic. A dagger marks configurations with fewer than ten positives under at least one label set; these comparisons are excluded from significance tests.

Combination	TDE	Pnorm	Last-step	Completion-only
DS-7B/MATH†	.881→.533	.978→.650	.938→.531	.974→.567
Qwen2.5/MATH†	1.00→.461	1.00→.666	1.00→.491	.944→.522
Llama/MATH†	.804→.484	.830→.611	.959→.572	.800→.544
DS-7B/GSM8K	.830→.524	.878→.496	.892→.568	.665→.474
DS-7B/HotpotQA	.594→.497	.642→.604	.664→.505	.536→.519
Mistral/MATH†	1.00→.776	.909→.923	.970→.556	.768→.770
Gemma/MATH	.978→.775	.922→.719	.987→.734	.742→.533
Qwen2.5/GSM8K	.904→.679	.899→.749	.833→.598	.806→.587
Gemma/HotpotQA	.715→.686	.661→.625	.662→.577	.620→.600
Llama/HotpotQA	.721→.648	.762→.752	.694→.657	.606→.575
Mistral/GSM8K	.818→.723	.822→.756	.850→.726	.727→.609
Llama/GSM8K	.917→.621	.873→.677	.905→.657	.802→.619
Qwen2.5/HotpotQA	.783→.792	.826→.813	.678→.669	.538→.541
Gemma/GSM8K	.980→.808	.954→.886	.979→.795	.540→.432
Mistral/HotpotQA	.549→.568	.604→.604	.654→.660	.516→.515

The effect extends beyond the core models but is not uniform. All nine tested decreases on DS-32B and Qwen3 remain significant after correction over twelve planned comparisons. Qwen3/MATH has no short-budget positives, so its short-budget AUC is undefined (Appendix D.1). ThinkPRM-1.5B also depends on the target, but with mixed direction: AUC falls from 0.737 to 0.589 on DS-7B/GSM8K and rises from 0.631 to 0.664 on Llama/MATH (Appendix D.2).

Reference-label training yields limited, variable recovery. The fixed five-fold control separates training-target and evaluation-target effects (Appendix E.1). On DS-7B/GSM8K, reference-trained reference AUCs are 0.456 for TDE, 0.521 for Pnorm, and 0.514 for the last-step probe; some configurations retain more signal, with Pnorm and last-step reaching 0.693 and 0.698 on DS-7B/MATH. These finite-sample results separate target correction from predictive capability without establishing an information limit for the prefix.

A larger Qwen2.5/GSM8K control uses 800 reference-labeled trajectories and obtains last-step AUC 0.584 against a short-budget-trained baseline of 0.621, showing no improvement at that scale. Three-state training, layer scans, prefix groups, and a permuted-label control provide further diagnostics in Appendices E.2–E.5.

5.3 RQ3: What accounts for the target sensitivity?

Completion cues explain part of the short-budget discrimination. The completion-only baseline changes from 0.974 to 0.567 on DS-7B/MATH and from 0.665 to 0.474 on DS-7B/GSM8K. Output form can therefore predict short-budget labels while providing little reference discrimination, though the baseline does not match every probe and completion-related confounding is broader than a single engine cap flag (Appendix F.1).

Reference-positive examples can receive low short-budget-trained scores. On DS-7B/GSM8K, the 136 examples positive only under the reference receive mean TDE score 0.07, compared with 0.48 for the 28 examples positive under both targets. Similar separations occur on Qwen2.5 and Llama GSM8K (Appendix C.3). The score ordering preserves the short-budget distinction among examples sharing a positive reference outcome.

Answer availability is a related signal. A gold-answer-presence oracle checks whether the correct answer occurs verbatim in the short output; its reference AUC overlaps the monitor range on several configurations and exceeds it on several HotpotQA rows (Appendix F.3). This is a privileged-information diagnostic: an occurrence can be intermediate or incidental, and the gold answer is unavailable at deployment.

The representation-level mechanism remains open. Removing one fitted completion direction changes step-level reference AUC by at most 0.015 (Appendix F.4); completion information may remain in other directions. Early-answer and elaboration instructions also change scores, but can change reasoning and correctness (Appendix G.2). The controls thus support completion-related confounding without identifying a unique causal feature.

6 Discussion

Define the outcome horizon explicitly. The observation budget, outcome budget, and handling of unfinished outputs should be stated separately. A within-budget failure is a valid negative for a within-budget objective, whereas a claim about a larger-budget outcome needs an aligned reference.

Audit before interpreting AUC. A paired label comparison can reveal target sensitivity without retraining: both directions of label change, class counts, and reference status belong alongside AUC.

Preserve unresolved outcomes. More tokens do not guarantee a definitive label; native caps, completion heuristics, answer markers, and prefix matches should remain separate metadata. A three-state representation can distinguish correct, incorrect, and unresolved outputs under a declared completion rule.

Make comparisons reproducible. The planned release, RM-RES, contains the 19 configurations and 2,886 paired trajectories with both label targets, monitor scores, completion indicators, and reference-status metadata. Aggregates should state their membership, use the same configurations under both targets, and exclude undefined AUCs.

Scope and limitations. The study uses greedy decoding, selected open models, specific monitors, and bounded references. Extraction fallbacks, imperfect completion indicators, and numerical variation limit attribution of individual changes to semantic truncation. Some references remain incomplete at larger budgets, and the safety prompts do not constitute a validated safety-monitor benchmark. These limits constrain generalization while leaving the observed sensitivity to label construction directly measurable.

7 Conclusion

Generation budgets help define the correctness target in reasoning-monitor evaluation. In our paired audit, replacing 256-token labels with operational 2048-token references changes trajectory judgments, model point-estimate rankings, and monitor AUC. Completion-related controls are consistent with the short-budget discrimination resting partly on completion and answer-availability cues, while retraining separates target alignment from predictive capability. The practical requirement is to audit the target before interpreting monitor performance: failing to finish within a short budget is not automatically failing under the outcome horizon the monitor claims to predict.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure: designing and giving feedback on parts of the methodology and the supplementary experiments (the reference-budget escalation study, the prior-work protocol audit, and the relation between label disagreement and AUC change); implementing methods (the analysis, extraction, and regeneration code used in the experiments); and formulating mathematical claims (the label-agreement formalization and the general probability relations in Appendix C.1). We have not used generative AI tools to generate synthetic datasets, to provide ingredients or writing for proofs, or to conduct qualitative or thematic analysis, and translation is not applicable to this work. Additionally, we used generative AI tools to identify and verify relevant literature and bibliographic information, and for drafting and editing portions of the paper text and for proofreading. We have reviewed all AI-assisted work: experiment designs and their reported values were checked against the underlying data by the authors, and the code was executed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The study uses public model checkpoints and datasets. Main comparisons use greedy bf16 generation with fixed prompts and extraction procedures, and prompts and answer extractors are held fixed across the 256-, 2048-, 4096-, and 8192-token budgets so that each comparison isolates the generation budget. Paired-prefix and larger-budget checks assess numerical reproducibility and reference stability. Appendix A.1 describes the available protocol; the remaining appendices identify the subsets, targets, and controls. Bootstrap analyses use 2,000 resamples with seed 42, and cross-validation keeps all observations from a trajectory in one fold. RM-RES comprises paired labels, scores, completion indicators, and reference status. A repository containing the analysis scripts, paired labels, monitor scores, and configuration-level reference status needed to reproduce the reported tables and figures will be made publicly available.

REFERENCES

Bowen Baker, Joost Huizinga, Leo Gao, et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yik Siu Chan, Zheng-Xin Yong, and Stephen H Bach. Can we predict alignment before models finish thinking? Towards monitoring misaligned reasoning models. *arXiv preprint arXiv:2507.12428*, 2025.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Qiguang Chen, Libo Qin, Jinhao Liu, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69:161101, 2026.

Maciej Chrabąszcz, Aleksander Szymczyk, Marcin Sendera, et al. Monitoring the internal monologue: Probe trajectories reveal reasoning dynamics. *arXiv preprint arXiv:2605.18549*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Joey David. Temporal predictors of outcome in reasoning language models. *arXiv preprint arXiv:2511.14773*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Keyu Duan, Zichen Liu, Xin Mao, et al. Efficient process reward model training via active learning. *arXiv preprint arXiv:2504.10559*, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Yichao Fu et al. Efficiently scaling LLM reasoning with Certaindex. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, et al. Does thinking more always help? mirage of test-time scaling in reasoning models. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tingxu Han et al. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset. In *Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2733–2743, 2019.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Daniel Kaiser, Arnoldo Frigessi, Ali Ramezani-Kebrya, and Benjamin Ricaud. Beyond accuracy: Decomposing the reasoning efficiency of LLMs. *arXiv preprint arXiv:2602.09805*, 2026.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, et al. Process reward models that think. *Transactions on Machine Learning Research (TMLR)*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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 (ICLR)*, 2024.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. Learn to reason efficiently with adaptive length-based reward shaping. In *International Conference on Learning Representations (ICLR)*, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, et al. s1: Simple test-time scaling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Wenhua Nie, Junlin Liu, Jianan Wu, et al. The coupling tax: How shared token budgets undermine visible chain-of-thought under fixed output limits. *arXiv preprint arXiv:2605.07686*, 2026.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Kai Ruan, Zihe Huang, Ziqi Zhou, et al. Doomed from the start: Early abort of LLM agent episodes via a recall-controlled probe cascade. *arXiv preprint arXiv:2607.06503*, 2026.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, et al. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Yucheng Sun, Alessandro Stolfo, and Mrinmaya Sachan. Probing for arithmetic errors in language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9426–9439, 2024.

Menghua Wu, Cai Zhou, Stephen Bates, and Tommi Jaakkola. Thought calibration: Efficient and confident test-time scaling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Bin Xie, Bingbing Xu, Xueyun Tian, et al. Towards robust process reward modeling via noise-aware learning. *arXiv preprint arXiv:2601.12748*, 2026.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Jingyang Yi, Jiazheng Wang, and Sida Li. ShorterBetter: Guiding reasoning models to find optimal inference length for efficient reasoning. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, et al. Hidden error awareness in chain-of-thought reasoning: The signal is diagnostic, not causal. *arXiv preprint arXiv:2605.09502*, 2026.

Anqi Zhang, Yulin Chen, Jane Pan, et al. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025a.

Yao Zhang, Yu Wu, Haowei Zhang, et al. GroundedPRM: Tree-guided and fidelity-aware process reward modeling for step-level reasoning. *arXiv preprint arXiv:2510.14942*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

APPENDIX

DS-7B and DS-32B denote the two DeepSeek-R1-Distill-Qwen models. Other abbreviations identify the checkpoints in Section 4. S and R denote 256-token and operational 2048-token labels unless specified otherwise. Results from different subsets or training procedures are reported separately.

A EVALUATION PROTOCOLS

A.1 PROMPTS, EXTRACTION, AND STATUS INDICATORS

The base instruction is:

> Solve the following task step by step. Use the exact format: Step 1: ... Step 2: ... ... Final Answer: ...

The dataset item follows, introduced by `Question:` for GSM8K and HotpotQA or `Problem:` for MATH. HotpotQA includes its provided context. Prompts remain unchanged within a budget comparison.

Extraction first checks the segment following `Final Answer:`, then alternative answer patterns such as `the answer is` and boxed expressions, with dataset-specific handling. GSM8K extracts the last numerical answer in the selected segment; MATH uses the last boxed expression; HotpotQA applies normalized entity matching. Without an explicit answer, fallbacks may inspect the full output or final line. These fallbacks can return intermediate values. Exact reproduction requires the executable normalizer, matcher, and fallback precedence in addition to this overview.

The text heuristic examines the last reasoning step for an operator, comma, open parenthesis, or missing sentence-final punctuation without a recognized answer marker. Markers include `Final Answer`, `The answer is`, and boxed-answer notation; punctuation checks include English and CJK endings. Main truncation estimates use 200 examples per configuration and seed 42. This rule is not a semantic annotation.

Native termination records an engine stop without a cap. Exact-prefix continuation instead tests whether a capped short output is a strict token prefix of its paired longer output. Numerical settings and attention implementations can change the greedy path, so these indicators are retained separately. The reference marker rate records explicit `Final Answer` presence; a marker-free output can still have a correct extracted answer.

TDE encodes top-layer displacements $h_{k+1} - h_k$ with a bidirectional LSTM and attention pooling, using a 64-dimensional representation, contrastive training, and a logistic head. Pnorm fits logistic regression to standardized step-final states and averages step probabilities. The last-step probe fits logistic regression to the final observed step's top-level state. These implementations assess the stated representations rather than replicating every prior probing protocol.

The Mistral checkpoint is `mistralai/Mistral-7B-Instruct-v0.3`. Checkpoint and tokenizer identity, precision, attention implementation, stopping settings, extraction code, and split identifiers are relevant reproducibility details. Prompt identity alone does not certify a shared trajectory.

A.2 LABEL CONSTRUCTION IN RELATED STUDIES

Table 5 distinguishes outcomes extracted from self-generation from targets supplied independently of the observed prefix. A reported cap identifies a protocol choice to inspect, not evidence that a published conclusion is invalid. An unreported cap does not imply unrestricted generation.

Table 5: Label-construction protocols in fourteen related studies. The table records the reported generation limit, label source, and handling of incomplete outputs. A identifies externally supervised or controlled targets; B identifies outcomes labeled from self-generated answers; C identifies prediction from prefixes with separately determined episode or trajectory outcomes. These categories describe protocol design and do not establish contamination of a published result. Budget-forcing studies are marked separately because they explicitly elicit a final answer after limiting reasoning.

Paper	Generation limit and cap handling	Label source	Protocol category
(Zhang et al., 2025a)	30K cap, over-cap discarded	Externally judged intermediate answers	A
(Sun et al., 2025)	Not specified in the audited description	Controlled arithmetic targets	A
(Yuan et al., 2026)	512 (CoT)/1024 (full); no cap handling described	Self-generated answer vs. reference string	B
(Baker et al., 2025)	Not specified in the audited description	Environment/regex ground truth	A
(Chan et al., 2025)	Thinking budgets 0.5–8K, then complete	External evaluator + human gold	C
(Ruan et al., 2026)	Episode round caps (10–20)	Episode outcome from the environment	C
(Chrabąszcz et al., 2026)	Not specified in the audited description	Math: self-generated answer, exact match; safety: template labels	B/A
(Lightman et al., 2024)	Not specified in the audited description	800K human step labels	A
(Wang et al., 2024)	Not specified in the audited description	Rollout final answers vs. golden answer	B
(Khalifa et al., 2026)	8K cap on verification generation; force-decode beyond	PRM800K human step labels (gold-matched)	A
(Muennighoff et al., 2025)	Budget forcing, forced answer	Answer after forced exit vs. reference	B‡
(Ghosal et al., 2025)	Exact budgets 256–16K, forced answer	Forced final response vs. dataset gold	B‡
(David, 2025)	Greedy, 512-token cap; no cap handling described	Eventual-answer correctness	B
(Han et al., 2025)	Prompt-specified budget, no hard cap	Final answer vs. dataset gold	–

‡ Deliberate budget forcing that always produces a final answer—a different mechanism from mid-reasoning truncation; not monitor papers.

B Budget Dependence and Reference Stability

B.1 Full budget sweep

The sweep covers DS-7B on MATH and GSM8K, Llama on MATH and GSM8K, and DS-32B and Qwen3 on GSM8K. Each configuration uses 50 held-out prompts at 128, 256, 512, 1024, 2048, and 4096 tokens.

The figure reports heuristic truncation and agreement with the operational 2048-token reference. Reusing the same labels at the 2048 point gives agreement one by construction; a separately generated 2048 reference can instead expose rerun differences. The pairing procedure therefore matters. Fixed-short-score and extended-output AUC comparisons are reported separately in Table 8.

Additional intermediate-budget measurements on reported 50-example DS-7B subsets give positive-label rates of 24% / 64% / 70% on GSM8K and 12% / 30% / 38% on MATH at 512 / 1024 / 2048 tokens. A separate completion check reports 27.5% / 84.5% on GSM8K and 7.5% / 29.0% on MATH at 512 / 1024 tokens. These series describe different analyses and are not combined into one paired estimate.

B.2 Stability at 4096 and 8192 tokens

Table 6 compares the primary reference with 4096-token labels on 50 paired examples per configuration. Several configurations are stable, while DS-7B/MATH and HotpotQA remain sensitive. Changes in both directions preclude interpreting the 2048 label as a per-example lower bound on later correctness.

We extend the two low-completion DS-7B configurations to 8192 tokens. On MATH, the completion indicator rises from 52% at 4096 to only 54% at 8192. Agreement is 74% between 2048 and 8192 labels and 66% between 4096 and 8192 labels. Fixed-score last-step AUC is 0.576 / 0.527 / 0.462

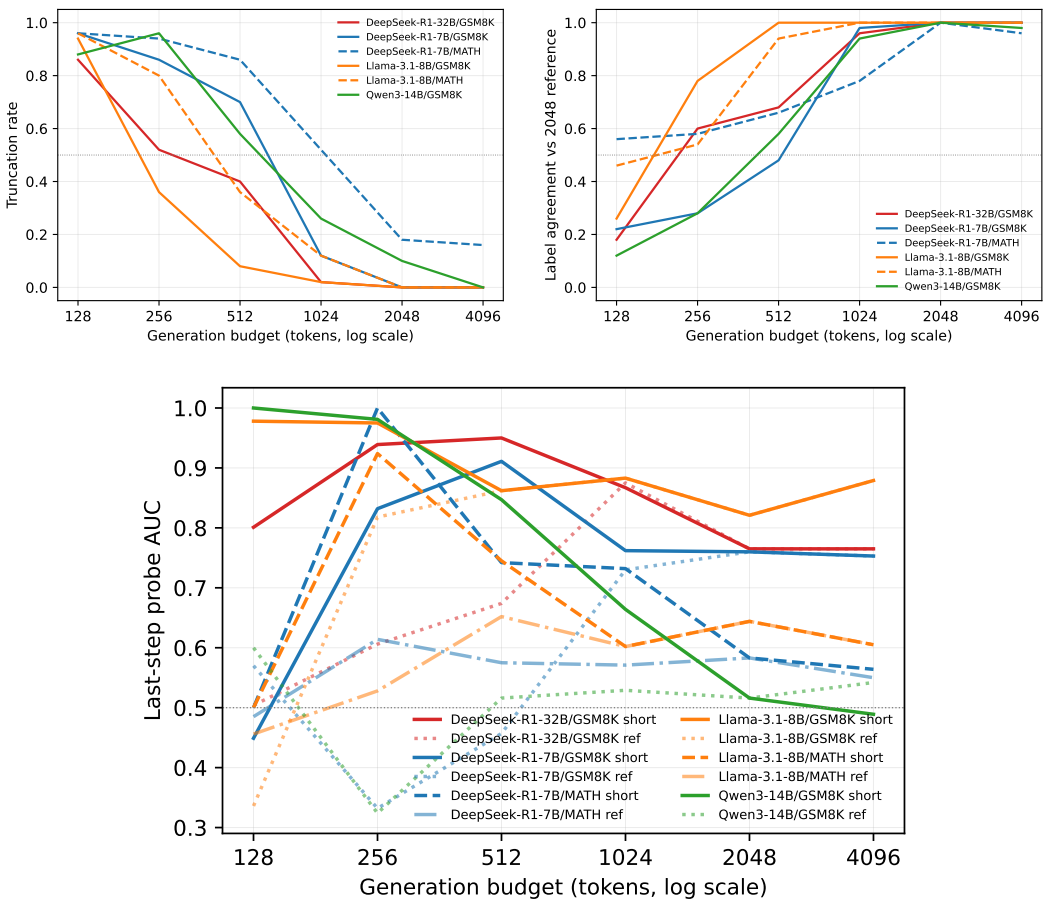

Figure 3: Budget sensitivity across all six configurations. (a) Heuristic truncation rate. (b) Agreement with the 2048-token reference labels. (c) Last-step probe AUC evaluated against budget-specific and reference labels. Each configuration uses 50 held-out examples. The panel legends identify model–dataset pairs and label targets.

Table 6: Stability of the 2048-token reference under a 4096-token budget. Each row uses 50 paired examples. Label changes are counted from 2048 to 4096 tokens. Agreement and answer-marker rates are percentages; the two change columns report example counts.

Combination	Agreement (%)	wrong→right / right→wrong	Completion 2048 / 4096 (%)
DS-7B/MATH	76	7 / 5	40 / 52
DS-7B/GSM8K	96	1 / 1	90 / 96
DS-7B/HotpotQA	62	14 / 5	38 / 76
Llama/GSM8K	90	4 / 1	92 / 98
Qwen2.5/GSM8K	92	2 / 2	100 / 100
Mistral/GSM8K	96	2 / 0	100 / 100
Gemma/GSM8K	100	0 / 0	100 / 100
DS-32B/MATH	96	2 / 0	90 / 96
DS-32B/GSM8K	100	0 / 0	98 / 98
Qwen3/MATH	96	2 / 0	86 / 96
Qwen3/GSM8K	98	1 / 0	96 / 100

under the 2048 / 4096 / 8192 references; TDE is 0.626 / 0.588 / 0.454 and Pnorm 0.720 / 0.618 / 0.582. These are sensitivity measurements against an unresolved target.

On HotpotQA, the completion indicator remains 76% at 4096 and 8192, whose labels agree on 90% of examples. Last-step AUC is 0.578 / 0.578 / 0.627 across the three references, and Pnorm is 0.620 / 0.620 / 0.783. This configuration remains flagged but does not show MATH's pattern. Operational labels are retained alongside reference status.

B.3 Is the reference a continuation of the cached prefix?

A paired regeneration is not automatically a continuation of the cached prefix. Greedy runs can diverge under numerical or implementation differences, so the 2048-token output need not be the continuation of the particular 256-token prefix whose monitor score is being evaluated. Where the cached output is a strict token prefix of its paired regeneration, the two runs demonstrably followed the same path up to the cap, and the reference is a genuine continuation of the observed prefix. This condition requires both that the short run reached its cap rather than stopping naturally and that the paired run reproduced it token for token; it therefore measures how often the two runs coincide below the cap, not how often they agree.

Table 7: Same-path continuation subset. n is the full reference subset and n_{path} the examples whose cached 256-token output is a strict token prefix of its paired 2048-token output. Agreement is reported over all examples and over that subset. Monitor entries are short-label AUC → reference-label AUC *restricted to the same-path subset*; all-example values are in Table 4. A subset is selected on numerical reproducibility rather than sampled, so configurations with few same-path examples describe a small selected group.

Configuration	n	n_{path}	Agree.	Last-step	TDE	Pnorm
DS-7B/GSM8K	200	186	0.32→0.28	0.856→0.568	0.804→0.533	0.865→0.493
DS-7B/HotpotQA	150	127	0.77→0.77	0.644→0.501	0.587→0.513	0.632→0.619
DS-7B/MATH	140	137	0.67→0.66	0.750→0.500	0.522→0.502	0.926→0.626
Gemma/GSM8K	200	34	0.92→0.53	0.989→0.491	0.957→0.463	0.935→0.702
Gemma/HotpotQA	100	19	0.92→0.84	0.538→0.452	0.564→0.560	0.603→0.369
Gemma/MATH	100	57	0.91→0.86	1.000→0.520	1.000→0.726	0.964→0.520
Llama/GSM8K	200	199	0.71→0.71	0.904→0.654	0.926→0.635	0.875→0.681
Llama/HotpotQA	100	99	0.90→0.90	0.696→0.659	0.719→0.645	0.757→0.747
Llama/MATH	96	96	0.62→0.62	0.959→0.572	0.804→0.484	0.830→0.611
Mistral/GSM8K	200	72	0.88→0.76	0.813→0.491	0.846→0.604	0.856→0.653
Mistral/MATH	100	62	0.99→0.98	NA→0.230	NA→0.836	NA→0.967
Qwen2.5/GSM8K	200	137	0.78→0.71	0.824→0.557	0.909→0.643	0.889→0.686
Qwen2.5/HotpotQA	100	26	0.95→0.85	0.650→0.611	0.408→0.556	0.750→0.778
Qwen2.5/MATH	100	95	0.64→0.63	NA→0.482	NA→0.440	NA→0.660

Table 7 restricts the audit to those same-path examples. On configurations where such examples are common, the short-budget-versus-reference gap is preserved. On DS-7B/GSM8K, the last-step probe's reference AUC is 0.568 over all 200 examples and 0.568 over the 186 same-path examples, while the short-label AUC moves only from 0.892 to 0.856. Llama/MATH is identical on all 96 examples. The pattern holds for TDE and Pnorm as well. Where the same-path subset is small or its agreement shifts substantially, the estimates describe a selected subset rather than the full configuration. Gemma/GSM8K (34 of 200), Gemma/HotpotQA (19 of 100), and Qwen2.5/HotpotQA (26 of 100) fall in this category and enter the table for completeness. Mistral/HotpotQA has only 5 same-path examples and is omitted, because the minimum of ten positives and ten negatives that governs the other tables cannot be met there.

The evidence therefore applies to the same-path examples and says nothing about the unmatched remainder, where the cached prefix and its regeneration diverged. Nor does it establish what a forced continuation from the cached prefix would produce, which would require decoding from the prefix itself rather than re-running the prompt. Within that scope, the short-budget-versus-reference gap is not an artifact of divergent paired runs.

B.4 ADAPTIVE REFERENCE BUDGETS

The adaptive rule selects the smallest tested budget with at least 95% heuristic completion. Table 8 separates changing the target for a fixed 256-token score from scoring a newly extended output. Meeting the heuristic criterion does not guarantee semantic completion or label stability.

Table 8: Evaluation with an adaptive reference budget. For each 50-example subset, B_{ref} is the smallest tested budget with at least 95% heuristic completion. AUC(B_{ref}) uses scores from outputs generated at budget B_{ref} and evaluates them against labels y^*. A flagged row does not meet the completion criterion; its largest tested budget is shown only as a diagnostic comparison.

Combo	B_{ref}	AUC(256, y_{256})	AUC(256, $y_{B_{\mathrm{ref}}}$)	AUC(256, y_{2048})	AUC(B_{ref}, $y_{B_{\mathrm{ref}}}$)
DS-7B/GSM8K	2048	.832	.332	.332	.760
DS-7B/MATH†	4096	1.00	.615	.614	.564
Llama/GSM8K	1024	.975	.818	.818	.883
Llama/MATH	2048	.924	.528	.528	.644
DS-32B/GSM8K	1024	.939	.583	.606	.867
Qwen3/GSM8K	4096	.981	.337	.324	.489

Where the criterion is met, adaptive-reference AUC is close to the 2048-reference value. The final column answers a different question because the observed representation changes. Reference selection and the benefit of additional observed computation should therefore be interpreted separately.

B.5 EXPLORATORY AIME TRANSFER

The AIME analysis uses 133 problems from 2020–2024, with probes trained on each model's MATH data and no AIME training. Both models have zero short-budget positives and eight reference positives. Agreement is 94% despite a 100% cap rate.

Table 9: Exploratory transfer to AIME. Each model is evaluated on 133 problems from 2020–2024 using a probe trained on its MATH data. Both models have zero short-budget positives and eight reference positives. Short-budget AUC is therefore undefined; reference AUC estimates have limited precision.

Model	Cap@256	Acc@256	Acc@2048	Agree.	Probe AUC (short/ref)
DeepSeek-R1-7B	1.00	0.0%	6.0%	94%	NA / .324
Qwen3-14B	1.00	0.0%	6.0%	94%	NA / .460

This illustrates how high agreement can conceal a target difference when positives are rare. It does not estimate an AUC decrease, because the short-budget AUC does not exist.

C LABEL DECOMPOSITIONS

C.1 GENERAL RELATIONS

Table 10 reports the operational quantities used in the analytic comparison. On the reference subsets, heuristic truncation and exact-prefix cap events need not agree. In particular, a capped output may already contain an answer, while a textually complete step may still precede an unfinished solution. These distinctions explain why the measured agreement among heuristic-complete outputs can be below one.

The idealized relations in Section 3.2 assume that an extractor assigns no positive label to a semantically truncated output and that untruncated paired outputs retain the same label. The general expressions make the missing terms explicit. Define

$$q_+ = P(y_s{=}1 \mid E,\, y^*{=}1), \qquad q_- = P(y_s{=}1 \mid E,\, y^*{=}0), \tag{5}$$

and let $a = P(y_s{=}y^* \mid \neg E)$ and $u_+ = P(y_s{=}0,\, y^*{=}1 \mid \neg E)$. Total probability gives

$$P(y_s{=}y^*) = (1-\tau)\,a + \tau\big[p_E^* q_+ + (1-p_E^*)(1-q_-)\big], \tag{6}$$

and, for $p^* > 0$,

$$P(y_s{=}0 \mid y^*{=}1) = \frac{\tau\, p_E^*\,(1-q_+) + (1-\tau)\,u_+}{p^*}. \tag{7}$$

Equation (7) differs from Equation (2)'s second expression in two ways: the factor $(1-q_+)$ allows a truncated output to be labeled positive, and the additive term $(1-\tau)u_+$ accounts for label changes outside E. Setting $q_+ = 0$ removes the first; the second vanishes only if untruncated paired outputs never change label. Table 10 shows that a is often well below one, so this term is not negligible in practice. These identities do not require independence between fallback extraction and the reference outcome. Equation (2) follows when $a{=}1$, $q_+{=}q_-{=}0$, and $u_+{=}0$. For operational indicators, the corresponding quantities must be estimated on the same subset before comparing predictions with observations.

Table 10: Operational quantities used to assess the truncation model. τ_{sub} is the heuristic truncation rate on the reference subset; a is label agreement among outputs judged complete; and p_t^* is reference accuracy among outputs judged truncated. The residual is measured agreement minus $(1-\tau_{\text{sub}})a + \tau_{\text{sub}}(1-p_t^*)$, the general relation with the fallback terms set to zero; it is therefore not a test of the idealized Eq. (2). τ_{cap} is the exact-prefix cap rate, cut-after-answer is the conditional answer-present fraction among cap events, and τ_E is defined in Eq. (4).

Combination	τ_{sub}	a	p_t^*	Agreement residual	τ_{cap}	cut-after-answer	τ_E
DS-7B/MATH	.92	.46	.31	$+.00$	.98	.01	.97
Llama/MATH	.90	.40	.37	$-.02$	1.00	.06	.94
Qwen2.5/MATH	.87	.69	.37	.00	.95	.00	.95
Mistral/MATH	.53	.98	.00	.00	.62	.00	.62
Gemma/MATH	.43	.95	.14	.00	.57	.02	.56
DS-7B/GSM8K	.82	.58	.83	$-.07$	.93	.10	.84
Llama/GSM8K	.29	.87	.71	$-.02$	1.00	.59	.41
Qwen2.5/GSM8K	.35	.92	.57	$-.03$	.69	.43	.39
Mistral/GSM8K	.31	.93	.29	$-.02$	.36	.08	.33
Gemma/GSM8K	.10	.95	.47	$-.01$	.17	.09	.16
DS-7B/HotpotQA	.80	.70	.17	$+.04$	.85	.17	.70
Llama/HotpotQA	.25	.92	.32	$-.04$	.99	.44	.55
Qwen2.5/HotpotQA	.16	.96	.31	$-.03$	.26	.23	.20
Mistral/HotpotQA	.02	.96	.00	.00	.05	.20	.04
Gemma/HotpotQA	.32	.94	.25	$-.04$	.19	.32	.13

The reported residuals range from approximately -0.075 to $+0.040$. Equation (7) discrepancies range from -0.14 to $+0.08$ using the main truncation estimates and -0.11 to $+0.12$ using subset estimates, excluding Mistral/MATH. The interval based on the operational τ_E contains the observed agreement in all fifteen core configurations, but this empirical coverage is not a guarantee for an imperfect truncation indicator.

Agreement among heuristic-complete outputs can be well below one because a complete-looking step is not necessarily a finished solution, a fallback returns an intermediate answer, or a paired run diverges. Back-solved quantities cannot independently validate an agreement equation; the reported operational quantities also do not identify every term in Equations (6)–(7).

C.2 Composition of the short-budget negative class

Table 11 partitions short-budget negatives by reference marker presence and correctness. Marker-free outputs can contain either extracted label. Accordingly, the marker-present/correct fraction can be lower than the full negative-to-positive change rate. For Qwen3/GSM8K, 84.6% of short negatives have marker-present, positive references; for DS-7B/MATH, 61.8% lack the marker. The reference procedure therefore resolves different amounts of the requested output format across configurations.

C.3 Scores across paired-label outcomes

The four groups in Table 12 separate the two targets. On several GSM8K configurations, $(0,1)$ examples receive much lower mean scores than $(1,1)$ examples, although both are reference-positive.

Table 11: Composition of the short-budget negative class. The first column gives the fraction of 256-token negatives flagged as truncated. The next three columns partition those negatives by reference-side state: completed and correct, completed and incorrect, or incomplete. The final columns give last-step probe AUC for correctness among completed outputs and for incomplete versus completed outputs, respectively.

	Truncated	Reference states among negatives (%)			Probe AUC
	(256-side, %)	marker, correct	marker, incorrect	marker absent	present / absent
DS-7B/MATH	94.9	14.7	23.5	61.8	.505 / .408
Llama/MATH	93.3	20.0	33.3	46.7	.539 / .386
DS-7B/GSM8K	86.6	76.7	17.4	5.8	.539 / .369
DS-7B/HotpotQA	81.3	11.4	25.2	63.4	.479 / .427
Qwen2.5/GSM8K	67.7	39.8	59.1	1.1	.594 / .141
Mistral/GSM8K	49.2	15.8	82.5	1.7	.730 / .505
Llama/GSM8K	64.6	62.2	30.5	7.3	.654 / .379
Gemma/GSM8K	10.8	8.4	91.0	0.6	.795 / .513
DS-32B/MATH	80.9	61.0	30.5	8.5	.545 / .193
DS-32B/GSM8K	87.3	80.0	17.3	2.7	.628 / .112
Qwen3/MATH	92.5	62.5	20.5	17.0	.535 / .465
Qwen3/GSM8K	93.8	84.6	8.0	7.4	.501 / .282

This describes the ordering learned under the short-budget target without identifying the mechanism of an individual prediction.

Table 12: TDE scores by paired label outcome. A: positive under both labels; B: positive only at 256 tokens; C: positive only at 2048 tokens; D: negative under both labels. Each entry is the mean score with the number of examples in parentheses. A dash indicates an empty group.

Combination	A	B	C	D
DS-7B/MATH	.92 (3)	.01 (1)	.04 (45)	.02 (91)
Llama/MATH	.43 (4)	.90 (2)	.01 (34)	.03 (56)
DS-7B/GSM8K	.48 (28)	– (0)	.07 (136)	.16 (36)
DS-7B/HotpotQA	.25 (9)	.28 (18)	.23 (17)	.23 (106)
Qwen2.5/GSM8K	.82 (99)	.70 (8)	.06 (37)	.27 (56)
Mistral/GSM8K	.70 (74)	.43 (6)	.13 (19)	.27 (101)
Llama/GSM8K	.92 (113)	.95 (5)	.15 (52)	.37 (30)
Gemma/GSM8K	.80 (31)	.87 (3)	.02 (14)	.05 (152)

C.4 Agreement thresholds

The label-agreement warning threshold is descriptive. Thresholds of 60%, 70%, and 80% flag five, eight, and eleven configurations, respectively, using Tables 2 and 3. The sets are nested, but their membership changes with the threshold. We therefore report the continuous agreement values and use the threshold only to prioritize inspection.

The high-truncation group in Table 15 contains DS-7B/GSM8K and DS-7B/HotpotQA; the remaining group contains the other twelve eligible core configurations. This membership is defined on the core family only, and the truncation rates that define it come from the 200-example truncation samples rather than from the smaller reference subsets used for the AUC summary.

D Additional Monitor Comparisons

D.1 Additional models

All nine tested hidden-state AUC decreases remain significant after correction over twelve planned comparisons. Qwen3/MATH remains in the planned family but has no short-budget positives and cannot contribute a short-budget AUC.

Table 13: Fixed-score AUC comparisons for the additional models. Entries are AUC(y_B) → AUC(y^*), following Table 4. Qwen3/MATH has no positive short-budget labels, so its short-budget AUC is undefined and is excluded from testing.

Method	R1-32B/MATH	R1-32B/GSM8K	Qwen3/MATH	Qwen3/GSM8K
TDE	.941→.526	.932→.581	NA→.506	.869→.583
Pnorm	.959→.621	.933→.688	NA→.451	.848→.622
Last-step	.966→.572	.905→.644	NA→.571	.835→.554
Completion-only	.904→.480	.886→.588	NA→.454	.561→.450

D.2 THINKPRM-1.5B

The first two columns of Table 14 change only the evaluation target. The third also restricts the population to reference outputs with a final-answer marker. The mixed direction of the full-population comparison shows that sensitivity depends on the verifier–configuration pair.

Table 14: ThinkPRM-1.5B evaluated against the two label targets. Short and reference columns use the same trajectory scores. The final column evaluates reference-label AUC on the subset whose reference output contains the final-answer marker. Scores are obtained with the verifier's verification-chain procedure.

Combination	Short budget	Reference	Reference, completed
DS-7B/MATH	0.660	0.524	0.615
DS-7B/GSM8K	0.737	0.589	0.611
DS-7B/HotpotQA	0.481	0.564	0.552
Llama/MATH	0.631	0.664	0.754
Llama/GSM8K	0.836	0.694	0.733
Qwen2.5/GSM8K	0.634	0.601	0.601
Mistral/GSM8K	0.823	0.771	0.769
Gemma/GSM8K	0.777	0.582	0.581

D.3 AGGREGATE MEMBERSHIP

Eligibility requires at least ten positive and ten negative examples under both targets. The high-truncation group uses the Table 1 heuristic threshold of 74% and is evaluated over the core family only; configurations in the additional family enter the remaining group regardless of their truncation rate. ThinkPRM contributes only the core-family configurations on which it was evaluated. This screening-based grouping is descriptive because truncation and AUC estimates can use different subsets.

Table 15: Mean AUC across reference-audit configurations, recomputed from unrounded RM-RES scores. Values are macro-averages over configurations with at least ten positive and ten negative examples under both label targets; the same configurations enter the two label columns for each method. "High" groups the *core-family* configurations with $\tau \geq 74\%$ (Table 1); configurations in the additional family enter "Remaining" regardless of their truncation rate, so the split is not the same as applying the threshold to all included configurations. n is the number of included configurations. ThinkPRM-1.5B is averaged only over the configurations on which it was evaluated. Undefined AUCs are excluded.

	High truncation		Remaining	
Method	n	Short → reference	n	Short → reference
TDE	2	0.712 → 0.511	12	0.842 → 0.666
Pnorm	2	0.760 → 0.550	12	0.839 → 0.709
Last-step probe	2	0.778 → 0.537	12	0.829 → 0.653
ThinkPRM-1.5B	2	0.609 → 0.576	4	0.768 → 0.662
Completion-only	2	0.601 → 0.496	12	0.687 → 0.544

E Training Targets and Representations

E.1 Training-label and evaluation-label effects

Table 16 uses fixed five-fold splits. Each monitor is trained on short-budget or reference labels, and each resulting score vector is evaluated against both targets. These are pooled out-of-fold predictions, distinct from the original-training-set scores in Table 4. Differences between the tables therefore cannot be attributed only to relabeling. Class balance, training size, and reference quality also matter; Mistral/MATH has only two reference positives.

Table 16: Training-label and evaluation-label controls. Entries report pooled out-of-fold AUC from fixed five-fold splits, in the order short→short / short→reference / reference→reference / reference→short. The first term specifies the training target and the second the evaluation target. Configurations with fewer than ten positives under either label target are marked as having limited precision.

Combination	TDE	Pnorm	Last-step
DS-7B/MATH	.770/.501/.448/.816	.960‡/.574/.693/.921	.803‡/.544/.698/.908
Qwen2.5/MATH	.990/.461/.407/.974	1.00‡/.503/.583/.745	1.00‡/.547/.517/.964
Llama/MATH	.737/.602/.444/.607	.685‡/.635/.586/.706	.730‡/.471/.487/.361
DS-7B/GSM8K	.812/.475/.456/.327	.821/.525/.521/.560	.860/.490/.514/.473
DS-7B/HotpotQA	.454/.520/.432/.584	.508/.466/.457/.558	.495/.475/.453/.527
Mistral/MATH	.894/.605/.112/.131	.904/.615/.281/.293	.904/.467/.066/.141
Gemma/MATH	.914/.697/.638/.950	.711/.764/.763/.791	.875/.650/.651/.767
Qwen2.5/GSM8K	.882/.654/.638/.881	.791/.636/.643/.616	.809/.608/.509/.645
Gemma/HotpotQA	.691/.665/.660/.695	.523/.487/.450/.484	.602/.535/.544/.600
Llama/HotpotQA	.677/.590/.559/.629	.671/.627/.591/.611	.644/.561/.508/.561
Mistral/GSM8K	.795/.706/.683/.794	.647/.606/.584/.647	.805/.705/.643/.727
Llama/GSM8K	.905/.623/.560/.871	.783/.649/.535/.619	.893/.623/.546/.680
Qwen2.5/HotpotQA	.741/.750/.771/.763	.727/.765/.752/.721	.472/.471/.447/.447
Gemma/GSM8K	.958/.786/.800/.963	.918/.870/.894/.910	.940/.797/.886/.914
Mistral/HotpotQA	.489/.521/.509/.512	.541/.523/.494/.498	.594/.618/.614/.600

E.2 Three-state training

The three alternatives in Table 17 use the original binary target, binary training after excluding heuristic-incomplete outputs, and a three-class target separating correct, incorrect, and heuristic-incomplete states. The three-class score is the predicted probability of the correct state. All evaluations use reference labels.

Excluding incomplete-looking outputs raises AUC by 0.016–0.039 on the first four rows but reduces it on Qwen2.5, Mistral, and Llama GSM8K. The additional-model results also vary. Separating states improves target description without guaranteeing better prediction.

E.3 Selected-layer probes

On DS-7B/GSM8K, all six reference AUCs in Table 18 lie between 0.502 and 0.541 despite much higher short-budget values. Other configurations retain more discrimination: Pnorm reaches 0.689 on Llama/MATH and 0.750 on Qwen2.5/GSM8K. The scan therefore describes configuration-specific sensitivity, not an absence of useful information throughout the network.

E.4 Observed prefix-length groups

Table 19 averages last-step AUC over eight seeds. Each group contains observations assigned to the corresponding observed length. Differences can reflect both length and sample composition; they are not a controlled intervention that extends an identical prefix population.

Table 17: Reference-label AUC under alternative training targets. *Binary* uses the original short-budget label. *Completed-only* trains a binary classifier after excluding outputs judged incomplete. *Three-class* trains a multinomial classifier on correct, incorrect, and incomplete states and uses the predicted probability of the correct state. Δ is completed-only AUC minus binary AUC.

Configuration	Binary	Completed-only	Three-class	Δ
DS-7B/MATH	.497	.513	.548	+.016
Llama/MATH	.471	.487	.537	+.016
DS-7B/GSM8K	.527	.564	.509	+.037
DS-7B/HotpotQA	.526	.565	.551	+.039
Qwen2.5/GSM8K	.635	.510	.634	−.125
Mistral/GSM8K	.722	.610	.736	−.112
Llama/GSM8K	.635	.553	.636	−.082
Gemma/GSM8K	.770	.829	.794	+.059
DS-32B/MATH	.599	.656	.583	+.057
DS-32B/GSM8K	.621	.593	.621	−.028
Qwen3/MATH	.507	.535	.607	+.028
Qwen3/GSM8K	.466	.543	.478	+.077

Table 18: Probe performance at selected cached layers. Each entry is short-budget/reference AUC. Layers 0, 3, and 7 index the last eight cached transformer layers; layer 7 is the model's final layer. Probes are refit at each layer, so the final-layer entries are not the same fits as Table 4.

Combination	L0 Pnorm S/R	L3 Pnorm S/R	L7 Pnorm S/R	L0 Last S/R	L3 Last S/R	L7 Last S/R
DS-7B/MATH	.858/.579	.858/.576	.852/.582	.865/.553	.867/.579	.859/.584
Llama/MATH	.846/.644	.861/.689	.854/.616	.828/.526	.823/.525	.824/.514
DS-7B/GSM8K	.942/.502	.941/.510	.892/.522	.866/.534	.856/.541	.874/.528
DS-7B/HotpotQA	.657/.584	.695/.598	.682/.660	.584/.495	.673/.583	.602/.543
Qwen2.5/GSM8K	.877/.719	.902/.750	.796/.676	.893/.699	.898/.719	.813/.618
Mistral/GSM8K	.819/.766	.786/.714	.798/.734	.841/.754	.842/.754	.844/.748
Llama/GSM8K	.868/.685	.898/.732	.879/.729	.885/.631	.866/.649	.902/.663
Gemma/GSM8K	.959/.876	.947/.881	.938/.868	.962/.832	.965/.856	.962/.837

E.5 Permuted-label control

Monitors in Table 20 are trained on permuted reference labels and evaluated against the original labels on held-out examples in a five-fold design. The reported range is 0.379–0.642. A single permutation can give variable estimates on small samples; repeated permutations are needed to estimate a null distribution or establish a statistical conclusion.

E.6 Larger training set and numerical checks

With 800 reference-labeled Qwen2.5/GSM8K training trajectories, last-step reference AUC is 0.584 in the reported full evaluation and 0.536 in cross-validation, compared with a short-budget-trained baseline of 0.621. These procedures do not form a training-data scaling curve.

On 20-example checks per task, bf16/fp16 reference-label agreement is 95% on GSM8K, 70% on MATH, and 75% on HotpotQA. Matched bf16 step-final states have cosine similarity of at least 0.9999 to the original cache. Precision consistency reduces one source of variation but does not guarantee identical greedy paths across attention implementations. Native termination and paired-prefix records remain necessary.

F Completion and Answer-Availability Controls

F.1 Native termination and heuristic completion

Tables 21–22 compare native termination, the text heuristic, and the last-step probe on matched examples. Native termination can predict short-budget labels when both outcomes occur. On the

Table 19: Last-step probe AUC by observed prefix length. Entries give short-label AUC / reference-label AUC / reference-trained cross-validation AUC, averaged over eight seeds. Prefix lengths are measured in tokens from the 256-token trajectories.

Combination	0–64	64–128	128–192	192–224	224–256
DS-7B/MATH	.78/.61/.69	.81/.58/.66	.59/.53/.66	.57/.52/.61	.50/.50/.67
DS-7B/GSM8K	.72/.50/.54	.75/.53/.55	.78/.52/.52	.82/.51/.57	.80/.52/.50
DS-7B/HotpotQA	.56/.57/.55	.56/.53/.58	.50/.54/.50	.54/.58/.71	.60/.57/.58
Llama/MATH	.76/.62/.51	.73/.60/.64	.77/.62/.59	.98/.55/.49	.73/.55/.59
Llama/GSM8K	.63/.57/.60	.72/.63/.61	.75/.60/.59	.83/.60/.70	.89/.67/.51
Qwen2.5/GSM8K	.67/.61/.51	.70/.66/.56	.80/.70/.64	.74/.70/.65	.90/.60/.57
Mistral/GSM8K	.64/.62/.58	.64/.64/.61	.68/.64/.55	.75/.63/.56	.81/.62/.59
Gemma/GSM8K	.77/.70/.78	.89/.77/.76	.85/.66/.64	.87/.58/.45	.95/.53/.61

Table 20: Reference-label AUC after training with permuted targets. Monitors are trained with permuted reference labels in a five-fold design and evaluated against the original reference labels on held-out examples. The table reports the permutation run used in this control.

Combination	TDE	Pnorm	Last-step
DS-7B/MATH	.467	.439	.518
Llama/MATH	.440	.502	.516
DS-7B/GSM8K	.484	.542	.536
DS-7B/HotpotQA	.454	.430	.459
Qwen2.5/GSM8K	.642	.497	.461
Mistral/GSM8K	.379	.496	.534
Llama/GSM8K	.475	.552	.557
Gemma/GSM8K	.488	.504	.399

Llama sweep subsets it is effectively constant, yielding AUC 0.5 when both label classes exist, while the other scores vary. Cap status therefore does not capture every difference in answer availability within capped outputs.

Table 21: Native termination, heuristic completion, and probe scores on the budget-sweep subsets. Entries are short-budget → reference AUC at a 256-token generation budget, using the same 50 examples within each row. A constant indicator has AUC 0.5 when both label classes are present.

Combo	s_{native}	s_{comp} (heuristic)	Last-step probe
DS-7B/GSM8K	.739→.475	.696→.463	.832→.332
DS-7B/MATH	1.00→.523	.980→.568	1.00→.614
Llama/GSM8K	.500→.500	.908→.677	.975→.818
Llama/MATH	.500→.500	.935→.484	.924→.528
DS-32B/GSM8K	.918→.655	.920→.678	.939→.606
Qwen3/GSM8K	.556→.511	.611→.522	.981→.324

A separate joint-model analysis adds native termination to the probe score. Its reported mean reference-AUC change is $+0.012$ in the designated high-truncation group, ranging from -0.118 to $+0.125$, and $+0.043$ in the remaining group. This measures the increment from native termination given the probe. Measuring the reverse increment requires a joint-versus-native-only comparison. The grouping in this analysis should not be assumed identical to Table 15's recomputed membership.

F.2 Length and score–label associations

The length score in Table 23 is negative character count, so larger values indicate shorter outputs. Several feature AUCs and score–label correlations decrease under reference labels. Conditioning on reference marker presence changes the evaluated population and gives a separate diagnostic.

Table 22: Completion baselines and the last-step probe on matched reference subsets. Each row contains up to 100 examples, with the actual sample size shown in the final column. The cap rate uses the engine's native termination record. AUC entries compare the short-budget and reference targets on the same examples.

Combo	cap rate	s_{native}	s_{comp} (heuristic)	Last-step probe	n
DS-7B/MATH	.98	.833→.529	.974→.551	.914→.593	100
Qwen2.5/MATH	.96	.990→.531	.944→.522	1.00→.491	100
Llama/MATH	.99	.583→.513	.800→.544	.959→.572	96
DS-7B/GSM8K	.95	.604→.466	.535→.410	.823→.491	100
DS-7B/HotpotQA	.83	.525→.473	.527→.481	.688→.443	100
Mistral/MATH	.61	.808→.556	.768→.770	.970→.556	100
Gemma/MATH	.59	.831→.591	.742→.533	.987→.734	100
Qwen2.5/GSM8K	.73	.701→.612	.807→.544	.790→.518	100
Gemma/HotpotQA	.20	.591→.521	.620→.600	.662→.577	100
Llama/HotpotQA	.99	.511→.511	.606→.575	.694→.657	100
Mistral/GSM8K	.35	.740→.577	.747→.596	.851→.701	100
Llama/GSM8K	1.00	.500→.500	.817→.609	.905→.717	100
Qwen2.5/HotpotQA	.39	.597→.584	.538→.541	.678→.669	100
Gemma/GSM8K	.17	.599→.365	.558→.374	.978→.726	100
Mistral/HotpotQA	.03	.525→.523	.516→.515	.654→.660	100
Qwen3/GSM8K	.99	.528→.505	.593→.538	.827→.564	100
Qwen3/MATH	1.00	NA→NA	NA→.489	NA→.574	100
DS-32B/MATH	.60	.941→.562	.904→.518	.966→.636	100
DS-32B/GSM8K	.56	.928→.655	.940→.625	.926→.680	100

Table 23: Completion-related baselines and score–label associations. S and R denote short-budget and reference labels. Length uses the negative character count of the short output; completion uses the text-based heuristic. Correlations are Pearson correlations between monitor scores and labels. The final row reports TDE/Pnorm reference AUC on outputs with a reference-side final-answer marker.

	DS-7B/M	Llama/M	DS-7B/G	DS-7B/H	Qwen2.5/G	Mistral/G	Llama/G	Gemma/G
Length AUC S/R	.93/.58	.63/.52	.69/.48	.53/.35	.63/.63	.76/.67	.47/.40	.30/.20
Completion AUC S/R	.97/.57	.80/.54	.67/.47	.54/.52	.81/.59	.73/.61	.80/.62	.54/.43
corr(TDE) S/R	+.80/+.24	+.65/−.02	+.60/−.04	+.10/.00	+.75/+.32	+.53/+.38	+.79/+.20	+.83/+.60
corr(Pnorm) S/R	+.69/+.24	+.46/+.20	+.57/+.05	+.23/+.13	+.69/+.40	+.55/+.45	+.65/+.24	+.74/+.61
Marker subset	.67/.68	.50/.63	.54/.51	.53/.51	.68/.75	.72/.75	.60/.64	.81/.89

F.3 Gold-answer-presence oracle

The oracle checks whether the gold answer occurs verbatim in the short output. Its reference AUC overlaps the monitor range on several configurations and exceeds it on several HotpotQA rows. A gold answer can be an intermediate value or incidental mention. This privileged-information control therefore measures textual availability, not verified completion or a deployable capability.

F.4 Completion residualization

The residualization control removes a fitted completion direction from hidden states before step-level prediction. Absolute reference-AUC changes are at most 0.015. This shows insensitivity to that particular removal, while completion information can remain in other directions or correlated features.

G Score-Threshold Behavior and Additional Illustrations

G.1 Acceptance and calibration

Tables 26–27 use threshold 0.7 as an illustration, not a calibrated deployment rule. In Table 26, the completed-error pass rate is $P(s > 0.7 \mid y^*{=}0, \widehat{C}_{256}{=}1)$, where completion is the 256-side text heuristic. The pass-set error rate is $P(y^*{=}0 \mid s > 0.7)$. Their denominators differ. Reference-label

Table 24: Gold-answer-presence control. The binary oracle is one when the gold answer appears verbatim in the 256-token output. n is the number of examples in the reference subset and n^* the number of reference-positive examples. The table reports reference-label AUC, answer-presence frequency among reference-positive examples, and the AUC range of the three hidden-state monitors. The oracle uses the gold answer and is not a deployable baseline.

Combination	n	n^*	Oracle AUC	Gold answer present|correct	Probe ref-AUC range
DS-7B/MATH	140	48	.607	.312	.531–.650
Llama/MATH	96	38	.652	.632	.484–.611
Qwen2.5/MATH	100	38	.616	.474	.461–.666
Mistral/MATH	100	2	.908	1.000	.556–.923
Gemma/MATH	100	16	.738	.750	.719–.775
DS-7B/GSM8K	200	164	.567	.329	.496–.568
Llama/GSM8K	200	165	.643	.800	.621–.677
Qwen2.5/GSM8K	200	136	.612	.787	.598–.749
Mistral/GSM8K	200	93	.752	.785	.723–.756
Gemma/GSM8K	200	45	.763	.778	.795–.886
DS-7B/HotpotQA	150	26	.673	.692	.497–.604
Llama/HotpotQA	100	47	.873	.915	.648–.752
Qwen2.5/HotpotQA	100	44	.818	.886	.669–.813
Mistral/HotpotQA	100	35	.924	.971	.568–.660
Gemma/HotpotQA	100	40	.808	.700	.577–.686

Table 25: Step-level reference AUC before and after completion residualization. The residualized representation removes the fitted completion direction using the procedure described in Appendix F.4. The absolute AUC difference is at most 0.015 in the reported configurations.

Combination	Full AUC	Residual AUC
DS-7B/MATH	.574	.574
Llama/MATH	.584	.584
DS-7B/GSM8K	.512	.525
DS-7B/HotpotQA	.634	.638
Qwen2.5/GSM8K	.656	.641
Mistral/GSM8K	.633	.628
Llama/GSM8K	.644	.654
Gemma/GSM8K	.736	.726

ECE reaches 0.70 for the short-budget-trained DS-7B/GSM8K probe, illustrating that discrimination and calibration are separate properties.

G.2 Early-answer and elaboration instructions

Early-answer instructions are applied to examples with reference-negative original outputs. All three reported GSM8K mean changes are positive, but only Qwen2.5's interval excludes zero. Additional MATH checks do not show a consistent increase. The instruction may change the reasoning and answer as well as presentation.

Elaboration instructions are applied to 60 examples per configuration with reference-positive original outputs. The mean score decreases in both reported configurations, with an interval excluding zero for Qwen2.5. Post-intervention completion/accuracy rates are 66.7% / 66.7% for Llama and 61.7% / 50.0% for Qwen2.5. Since correctness changes too, the score shifts cannot isolate a purely superficial effect.

G.3 Safety-domain generation status

The defensive-knowledge analysis uses 60 prompts per model. Every 256-token output is flagged by the text heuristic, but exact-prefix cap rates range from 58.3% to 100%. The difference again distinguishes a textual indicator from an engine or semantic judgment.

Table 26: Score-threshold behavior of the last-step probe (short-budget-trained scores; threshold 0.7). *Completed-error pass rate* is the fraction of completed, reference-negative examples above the threshold, where completion is the 256-side text heuristic. *Pass-set error rate* is the fraction of accepted examples that are reference-negative. Parentheses give the relevant denominator; a dagger marks any cell whose relevant denominator is fewer than five, including the completed-error counts. ECE uses ten bins and reference labels (short-budget-trained / reference-trained). Empty denominators are reported as NA.

Combination	Completed-error pass (%)	Pass-set error (%)	ECE (S / R)
DS-7B/MATH	0.0 (3)†	0.0 (3)†	.32/.17
Llama/MATH	50.0 (4)†	66.7 (3)†	.40/.36
Qwen2.5/MATH	0.0 (7)	0.0 (2)†	.36/.36
DS-7B/GSM8K	12.5 (8)	7.7 (26)	.70/.18
DS-7B/HotpotQA	29.2 (24)	81.8 (22)	.25/.17
Llama/GSM8K	66.7 (18)	10.4 (125)	.33/.18
Qwen2.5/GSM8K	55.9 (34)	21.5 (107)	.33/.35
Mistral/GSM8K	30.2 (63)	27.3 (77)	.23/.26
Gemma/GSM8K	7.6 (145)	27.5 (40)	.12/.11
Llama/HotpotQA	41.7 (36)	40.4 (47)	.34/.36
Qwen2.5/HotpotQA	22.2 (45)	37.8 (37)	.20/.39
Mistral/HotpotQA	42.9 (63)	58.3 (48)	.35/.27
Gemma/HotpotQA	44.4 (36)	57.9 (38)	.43/.33

Table 27: Detailed score-threshold statistics (threshold 0.7). The first four columns compare short-budget-trained (S) and reference-trained (R) probes from the five-fold control on the same completed subset. The final two columns recompute the pass rates on the RM-RES records used in Table 26: reference-negative examples split by the 256-side text heuristic; parentheses give group sizes. Empty groups are reported as NA.

Combination	Error pass S	Error pass R	Pass-set err. S	Pass-set err. R	P(>0.7)|comp.	P(>0.7)|trunc.
DS-7B/MATH	.000	.094	.000	.300	.000 (3)	.000 (89)
Llama/MATH	.031	.156	1.000	.500	.500 (4)	.000 (54)
Qwen2.5/MATH	.000	.294	.000	.652	.000 (7)	.000 (55)
Mistral/MATH	.000	.000	–	–	.000 (45)	.000 (53)
Gemma/MATH	.044	.059	.500	.667	.064 (47)	.000 (37)
DS-7B/GSM8K	.067	.967	.118	.162	.125 (8)	.036 (28)
DS-7B/HotpotQA	.026	.000	1.000	–	.292 (24)	.110 (100)
Llama/GSM8K	.357	.893	.093	.146	.667 (18)	.059 (17)
Qwen2.5/GSM8K	.349	.587	.232	.311	.559 (34)	.133 (30)
Mistral/GSM8K	.200	.305	.318	.410	.302 (63)	.045 (44)
Gemma/GSM8K	.058	.071	.281	.289	.076 (145)	.000 (10)
Llama/HotpotQA	.485	.394	.457	.481	.417 (36)	.235 (17)
Qwen2.5/HotpotQA	.321	.321	.586	.630	.222 (45)	.364 (11)
Mistral/HotpotQA	.213	.197	.520	.600	.429 (63)	.500 (2)
Gemma/HotpotQA	.359	.256	.538	.417	.444 (36)	.250 (24)

Table 28: Probe-score changes after an early-answer instruction. The intervention is applied to examples whose original outputs are reference-negative. Before and after columns report mean last-step probe scores. Δ is the paired change, with a 95% bootstrap confidence interval.

Combination	n	Before	After	Δ [95% CI]
DS-7B/GSM8K	31	.129	.290	$+.161$ [$-.032$, $+.355$]
Llama/GSM8K	36	.493	.697	$+.203$ [$-.017$, $+.417$]
Qwen2.5/GSM8K	190	.304	.406	$+.102$ [$+.013$, $+.193$]

A separate set of 60 harmful-action prompts per model also has a 100% heuristic truncation rate at 256 tokens. At 2048, reported correct-refusal rates are 61.7% for Llama, 76.7% for Qwen2.5,

Table 29: Probe-score changes after an elaboration instruction. Each configuration uses 60 examples whose original outputs are reference-positive. Before and after columns report mean scores; Δ gives the paired difference with a 95% bootstrap confidence interval. The final column reports a two-sided test.

Combination	Before	After	Δ [95% CI]	p (two-sided)
Llama/GSM8K	.700	.567	$-.133$ [$-.275$, $+.009$]	.073
Qwen2.5/GSM8K	.701	.384	$-.317$ [$-.463$, $-.168$]	$<.001$

Table 30: Truncation indicators on defensive-knowledge prompts. Each model is evaluated on 60 prompts. The table reports heuristic truncation and exact-prefix cap rates at 256 tokens, and the completion indicator at 2048 tokens. These measurements concern generation status, not safety-monitor accuracy.

Model	256-token heuristic truncation	256-token exact cap	2048-token completion
Qwen2.5-7B	1.000	.583	1.000
Llama-3.1-8B	1.000	1.000	.417
Mistral-7B	1.000	.850	.933

and 5.0% for Mistral. Refusal assessment requires its own rubric and must be distinguished from answer correctness on the main tasks. These observations do not evaluate a trained safety monitor or establish its false-negative rate.