CURSE OF CORPUS: DO MODELS REASON BEYOND COUNTING AND COMPOSITION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Process reward models are often evaluated by pooling all labeled steps into one AUROC. We study this practice under the *absorbing-label convention*, which labels the first incorrect step and every later step as incorrect. Pooled AUROC then mixes two nuisances: comparisons across heterogeneous sources and an association between position and incorrectness. Both are measurable exactly, because a rank statistic is a sum over positive–negative pairs: on our held-out set only 0.59% of the 371,765,485 pairs compare steps of one source at one index, and 93.84% cross sources. A scorer that reads no text and knows only the corpus therefore reaches pooled absorbing AUROC 0.9140, against 0.9151 for a fine-tuned 31B verifier. Across five trained verifiers, held-out raw pooled AUROC exceeds mean-macro AUROC by 0.0963–0.1244. Across seven corpora, a text-blind relative-position baseline obtains absorbing-label pooled AUROC from 0.6370 to 0.9768; imposing absorption raises it by as much as 0.3124 without changing text or scores. We evaluate first-error detection with A_{hazard}, which excludes post-error suffixes and compares errors with correct steps at the same index. Absolute index alone therefore receives exactly 0.5, mean-macro and weighted-macro variants also exclude cross-source comparisons, and adding length strata returns relative position to 0.5. The estimands then disagree about direction and not only magnitude: averaged over the first and last five checkpoints of a matched 2,300-step run, every ranking estimand rises with an interval excluding zero, pooled AUROC by 0.0708 and weighted-macro A_{hazard} by 0.0363, while accuracy at naming the step that first broke falls by 0.0335, interval $[-0.0532, -0.0142]$. In a forced-verdict diagnostic, pooled raw AUROC places Qwen 0.0847 above Gemma whereas source-aware first-error intervals include zero; conversely weighted-macro A_{hazard} favors Muse over GLM by 0.0597 with interval $[0.0391, 0.0794]$ while pooled raw AUROC is inconclusive.

1 INTRODUCTION

A process reward model (PRM) scores intermediate steps in a reasoning trace. These scores support error diagnosis, solution ranking, and checkpoint selection. A common intrinsic evaluation pools all labeled steps and reports one AUROC. Although threshold-free, that number depends on both the label convention and the composition of the evaluation set.

We study the *absorbing-label convention*, under which the first incorrect step and every later step are labeled incorrect. The label then represents prefix validity rather than the correctness of the current step alone. Because later positions have had more opportunities to follow an error, a scorer can benefit from position without reading the reasoning; we call this *position leakage*. Our claim is specific to absorbing labels and does not directly extend to labeling schemes that permit recovery after an error.

Pooling heterogeneous sources adds a separate nuisance. Pooled AUROC both weights sources implicitly and compares their score scales directly. Macro aggregation removes cross-source comparisons but still permits comparisons between different positions. The central issue is therefore an estimand mismatch: invalid-prefix ranking is not the same task as identifying the step that first breaks a correct prefix.

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.

We evaluate the latter task with A_{hazard}. It retains each trace only through its first error and compares events with correct steps at the same absolute index. Mean-macro and weighted-macro variants make the same comparisons within each source, then weight sources equally or by eligible pair count. We also report incremental predictive information (IG), the held-out log-score gain from adding verifier scores to a baseline using only index and preceding token count.

Across corpus interventions, text-blind controls, five unadapted models, and five trained verifiers, the evaluation rule changes both rankings and which pairwise differences are supported. Pooled AUROC reports a large Qwen–Gemma gap that disappears under source-aware first-error metrics, while those metrics identify a Muse–GLM gap that pooled AUROC misses. On a matched training run every ranking estimand reports a resolved gain while accuracy at naming the first error reports a resolved loss, so the sign disagreement is between ranking steps and locating one rather than between two ranking rules.

We contribute (i) an exact decomposition of pooled step AUROC that measures how much of it is cross-source and cross-position comparison, (ii) a controlled measurement of position leakage with separate diagnostics for position and source identity, (iii) mean-macro, weighted-macro and length-stratified A_{hazard} estimators with invariance guarantees and a factor-by-factor ablation of each, and (iv) paired evidence that a ranking estimand and a first-error readout can reverse, and not merely blur, a conclusion about whether training helped.

2 Background

2.1 Process supervision and step labels

Outcome supervision grades a solution by its final answer (Cobbe et al., 2021). Process supervision grades intermediate steps so a verifier can identify where a solution fails (Uesato et al., 2022; Lightman et al., 2024). Existing benchmarks instantiate this objective differently. ProcessBench predicts the first-error location and reports the harmonic mean of accuracy on erroneous and correct traces (Zheng et al., 2025); PRMBench reports a composite score (Song et al., 2025); and Hard2Verify aggregates all step predictions across responses (Pandit et al., 2026). AUROC is one threshold-free statistic for such a pooled list (Xia et al., 2025).

Two labeling targets must be distinguished. A *local* label assesses a step itself and may return to correct after an earlier error. An *absorbing* label assesses whether the prefix remains valid. The latter matches the definition "whether the steps so far are correct" (Uesato et al., 2022) and the first-error-location strategy of Lu et al. (2024). PRM800K obtains the same shape by stopping annotation after the first negative step (Lightman et al., 2024). Monte Carlo labels can instead recover because they estimate each step from rollout success (Wang et al., 2024b). Neither target is universally correct: local labels measure step correctness, whereas absorbing labels measure prefix validity.

2.2 Shortcut diagnostics

Withholding an intended input is a standard test of a benchmark: hypothesis-only classifiers reveal artifacts in natural-language inference (Gururangan et al., 2018), reading-comprehension systems can answer without the passage (Kaushik & Lipton, 2018), and visual-question answering systems exploit priors without the image (Goyal et al., 2017), all of which reflects properties of datasets as well as models (Geirhos et al., 2020). In process supervision, deleting the natural-language content of a step can leave verifier scores nearly unchanged (Ma et al., 2025). Our controls go further by reading no text.

The closest empirical result is CROP, whose proxy combines surface counts, absolute position, normalized position, and trace length and reaches 0.992 step AUROC on PRM800K (Cheung et al., 2026). Because those inputs are combined, it does not isolate the source of the signal. We evaluate absolute position, relative position and source identity separately, and intervene on the labels themselves.

Statistically, comparing ranks only within groups that share a nuisance variable follows the van Elteren test (van Elteren, 1960) and covariate-adjusted ROC analysis (Janes & Pepe, 2008), while treating the first error as an event and retaining only traces that have not yet failed follows discrete-

time survival analysis (Cox, 1972; Allison, 1982) with an incident/dynamic ROC estimand (Heagerty & Zheng, 2005). We combine that risk set with exact-index stratification and Mann–Whitney pair weighting (Mann & Whitney, 1947; Bamber, 1975).

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let $t \in \{1, \dots, N\}$ index reasoning traces. Trace t has T_t steps, indexed by $i \in \{0, \dots, T_t - 1\}$. Let $c_{t,i} \in \{0, 1\}$ denote the dataset-native local correctness label. Define the first error and absorbing correctness label as

$$f_t = \min\big(\{i : c_{t,i} = 0\} \cup \{T_t\}\big), \tag{1}$$

$$a_{t,i} = \mathbb{1}[i < f_t]. \tag{2}$$

Here $f_t = T_t$ denotes a trace with no labeled error. The first error and every later step therefore receive $a_{t,i} = 0$.

Let $s_{t,i} \in \mathbb{R}$ be a verifier score, with larger values indicating greater evidence of error. For a chosen label vector $y_{t,i}$, pooled AUROC is the probability, with half credit for ties, that a randomly selected incorrect step receives a higher score than a randomly selected correct step. Its positive–negative comparison may cross both positions and sources.

3.2 MEAN- AND WEIGHTED-MACRO AUROC

Let A_d be the AUROC computed using only observations from source d, and let m_d^+ and m_d^- be its numbers of error-labeled and correct steps. For a fixed set $\mathcal{D}$ of sources containing both classes, define

$$A_{\text{AUROC}}^{\text{mean}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} A_d, \qquad A_{\text{AUROC}}^{\text{weighted}} = \frac{\sum_{d \in \mathcal{D}} m_d^+ m_d^- A_d}{\sum_{d \in \mathcal{D}} m_d^+ m_d^-}. \tag{3}$$

Mean-macro AUROC gives each source equal weight, weighted-macro AUROC gives each within-source positive–negative pair equal weight, and both exclude cross-source comparisons. Pooled AUROC remains a third, distinct estimand because it additionally contains comparisons between sources. We compute all three under raw and absorbing labels.

3.3 FIRST-ERROR RISK SET

For each trace, retain the steps before or at its first error, $\mathcal{R}_t = \{i : 0 \le i < T_t,\ i \le f_t\}$, and define the event indicator on it by $h_{t,i} = \mathbb{1}[i = f_t]$. An erroneous trace contributes its correct prefix as non-events and its first incorrect step as one event. A fully correct trace contributes all its steps as non-events. Post-error steps are excluded because neither absorbing prefix labels nor first-error annotations identify an additional local error there.

Let d_t denote the source of trace t and $\mathcal{R} = \{(t, i) : i \in \mathcal{R}_t\}$. Restricting to $\mathcal{R}$ changes the target from invalid-prefix detection to first-error detection, conditional on the preceding prefix being correct.

3.4 INDEX-STRATIFIED DISCRIMINATION

At index i, let $\mathcal{P}_i$ and $\mathcal{N}_i$ be the risk-set traces with $h_{t,i} = 1$ and $h_{t,i} = 0$, with sizes n_i^+ and n_i^-. For indices containing both classes, and with $\psi(a, b) = \mathbb{1}[a > b] + \frac{1}{2}\mathbb{1}[a = b]$,

$$A_i = \frac{1}{n_i^+ n_i^-} \sum_{u \in \mathcal{P}_i} \sum_{v \in \mathcal{N}_i} \psi(s_{u,i}, s_{v,i}), \qquad A_{\text{hazard}} = \frac{\sum_i n_i^+ n_i^- A_i}{\sum_i n_i^+ n_i^-}. \tag{4}$$

A_{hazard} is the average concordance over all event–non-event pairs sharing an absolute index. Because a scorer $s_{t,i} = g(i)$ is constant within every index, every such pair ties and $A_{\text{hazard}}(g(i)) = \frac{1}{2}$ exactly whenever the estimator is defined.

Global A_{hazard} may still compare two traces from different sources. To control source identity, define $A_{d,i}$ analogously within each source–index stratum, with counts $n^+_{d,i}$ and $n^-_{d,i}$, and write $A^{(d)}_{\text{hazard}} = \sum_i n^+_{d,i} n^-_{d,i} A_{d,i} / \sum_i n^+_{d,i} n^-_{d,i}$. We report two aggregations over these same comparisons,

$$A^{\text{mean}}_{\text{hazard}} = \frac{1}{|\mathcal{D}_H|} \sum_{d \in \mathcal{D}_H} A^{(d)}_{\text{hazard}}, \qquad A^{\text{weighted}}_{\text{hazard}} = \frac{\sum_d \sum_i n^+_{d,i} n^-_{d,i} A_{d,i}}{\sum_d \sum_i n^+_{d,i} n^-_{d,i}}, \tag{5}$$

the first giving each source equal weight and the second each comparable pair equal weight. $\mathcal{D}_H$ is fixed before evaluation, and if any expected source has no comparable first-error pair the mean-macro estimator is undefined rather than silently averaging over a smaller source set; we report the number of contributing sources and strata alongside it. Both source-aware variants give exactly 0.5 to a scorer that depends only on source and absolute index. Neither removes trace length or other features that vary within a source–index stratum; Section 5.3 adds a length stratum and measures what it removes and what it costs.

3.5 Incremental predictive information

We separately test whether verifier probabilities improve first-error prediction beyond structural information. Let $L_{t,i}$ be the number of input tokens preceding the prediction at step i, and let $z_{t,i} = (i, L_{t,i})$. On fitting observations, estimate the structural probability $\widehat{\pi}_{t,i} = \widehat{p}_{\text{str}}(h_{t,i} = 1 \mid z_{t,i})$ and the verifier-augmented probability $\widehat{p}_{t,i} = \sigma(\text{logit}(\widehat{\pi}_{t,i}) + q_\phi(s_{t,i}))$, where $\sigma(v) = (1 + e^{-v})^{-1}$ and the function class includes $q_\phi \equiv 0$.

For Bernoulli log score $b(h, p) = h \ln p + (1 - h) \ln(1 - p)$, and writing $\mathcal{R}_{\text{eval}}$ for the held-out part of the risk set $\mathcal{R}$, held-out incremental predictive information is

$$\text{IG} = \frac{1}{|\mathcal{R}_{\text{eval}}| \ln 2} \sum_{(t,i) \in \mathcal{R}_{\text{eval}}} \left[b(h_{t,i}, \widehat{p}_{t,i}) - b(h_{t,i}, \widehat{\pi}_{t,i}) \right]. \tag{6}$$

Positive IG means the verifier improves held-out log score, in bits per risk-set step. Mixed-source IG fits one pipeline across sources. Source-controlled IG fits the complete structural and augmented pipeline independently within each source; we report both observation-weighted and equal-source summaries. IG equals conditional mutual information only when the fitted predictors recover the true conditional distributions, so it does not by itself identify information caused by reasoning content. The IG's exact choices are detailed in Appendix C.

4 Experimental Setup

4.1 Data construction

The normalized training collection contains 25,039 reasoning traces, 207,490 supervised step examples, and 11,226 question groups from eight sources: ActPRM Duan et al. (2025), Double Delta Shen & Alonso (2026), Math-Shepherd Wang et al. (2024b), Med-PRM Yun et al. (2025), MMLU-Pro-CoT Wang et al. (2024c), PRM800K Lightman et al. (2024), PRM-Clinic Wang et al. (2024a) and Qwen ProcessBench Zheng et al. (2025). Double Delta receives separately constructed process supervision.

The corpus contains two judge-generated supervision streams produced using fine tuned configurations from the Qwen Qwen Team (2026) and DeepSeek DeepSeek-AI (2026) model families. Each record retains the judge rationale, reasoning steps, local gold label, absorbing label, and agent label. Judge rationales are conditioned on the labels supplied by the original corpus, not on an absorbing rewrite imposed for our evaluation intervention.

Each supervised example contains the question, any answer choices or source evidence, and the solution prefix through the step under review. It excludes future solution steps and all verifier rationales and verdicts from previous steps. Records are partitioned by (source, question identity) so that alternative solutions to the same question cannot cross splits.

The complete labeled test partition used for the corpus audit contains 55,524 traces and 384,253 steps from seven sources. Model evaluation uses a fixed held-out subset containing at most 1,000 traces

per source: 5,389 traces, 40,475 steps, and 1,427 question groups. This subset has no question-group overlap with training or validation.

Checkpoint evaluation uses a fixed source-balanced subset of 352 traces. The nominal budget of 400 yields 352 traces because 44 are taken from each of eight sources.

4.2 TEXT-BLIND CONTROLS

We use three controls. The *absolute-index* control assigns $s_{t,i} = i$. The *relative-position* control assigns $s_{t,i} = i/T_t$ and therefore additionally observes trace length. The *source-prior* control assigns every step from a source the same empirical first-error rate. We report it fit on the evaluation data, where it is an upper reference for what source identity alone can provide, and also fit on the validation split and applied unchanged to the held-out set, where it is deployable. The deployable version is the stronger of the two, reaching pooled absorbing AUROC 0.9228 on held-out against 0.9140 for the oracle-fit version, so the control is not an artifact of fitting on the test labels. None of the controls reads the question or reasoning text.

4.3 MODELS AND TRAINING

We evaluate five model families: Qwen3.6-27B (Qwen Team, 2026), Gemma 4 31B (Gemma Team, 2026), gpt-oss-20b (OpenAI, 2025), GLM-4.7-Flash (GLM Team et al., 2025), and Muse Glimmer 30B (Meta Superintelligence Lab, 2026). The unadapted diagnostic uses each released base model without an adapter and the trained experiment uses one LoRA run from each family. A separate public-PRM comparison evaluates Math-Shepherd 7B (Wang et al., 2024b), Qwen2.5-Math-PRM-7B (Zhang et al., 2025), and Qwen2.5-Math-7B-PRM800K (Zheng et al., 2025) through their native scoring interfaces.

Every trained model uses LoRA rank 32, scaling parameter 64, dropout 0.05, and peak learning rate 2×10^{-4}. Gemma and gpt-oss follow a matched 2,300-step schedule. Qwen, GLM, and Muse use architecture-specific batch configurations and training horizons, so we treat them as replications rather than members of the matched comparison. Appendix D gives the remaining settings.

4.4 SCORING PROTOCOLS

Trained checkpoints use generated-rationale scoring. The model greedily writes a short verification rationale followed by a `Verdict:` marker. At that position, we normalize the complete-sequence likelihoods of `VALID` and `INVALID` to obtain p_{invalid}. Missing verdicts are not imputed, and selection requires at least 0.99 coverage in every source. Appendix B gives the decoding settings.

The unadapted diagnostic instead appends `Verdict:` to the question and solution prefix and evaluates the two verdict likelihoods without generating a rationale. This forced-verdict protocol gives every model complete coverage on the same 15,301 steps. Because the score extraction and model adaptation differ, we do not interpret the unadapted and trained cohorts as a causal before–after comparison.

The public PRMs retain their native readouts: tag-token probabilities for Math-Shepherd and dedicated two-class heads for the Qwen PRMs. Their comparison is reported separately from both the forced-verdict diagnostic and the trained generated-rationale evaluation.

4.5 CHECKPOINT SELECTION

All five trained models are selected by the same rule: maximize raw-label mean-macro AUROC on the common 352-trace validation subset, subject to the 0.99 per-source coverage floor. An exact tie is resolved in favor of the earlier checkpoint. This selects Qwen step 350, Gemma step 1,600, gpt-oss step 2,200, Muse step 8,200, and GLM step 800. Selection uses validation labels only; held-out scores do not enter the rule. Appendix E reports candidate counts and validation values, and Appendix I gives the resampling design behind every interval in this paper. Held-out intervals condition on these fixed selections and do not include selection uncertainty.

Table 1: Five unadapted models and three text-blind controls on the same 15,301 steps. IG is source-controlled and reported in bits per risk-set step.

Type	Scorer	Raw pooled AUROC	Raw mean-macro AUROC	Global A_{hazard}	Mean-macro A_{hazard}	Weighted-macro A_{hazard}	Source-controlled IG
Model	Qwen3.6-27B	0.7653	0.7618	0.8770	0.8082	0.8504	0.0964
Model	Gemma 4 31B	0.6806	0.7768	0.8704	0.7987	0.8441	0.0975
Model	Muse Glimmer 30B	0.6689	0.7745	0.8214	0.7960	0.8057	0.0601
Model	GLM-4.7-Flash	0.6663	0.7072	0.8067	0.7191	0.7453	0.0374
Model	gpt-oss-20b	0.5314	0.5839	0.5993	0.6341	0.6162	0.0091
Control	Absolute index	0.7534	0.6644	0.5000	0.5000	0.5000	-0.0001
Control	Relative position	0.6368	0.7409	0.6187	0.6498	0.6621	0.0127
Control	Source prior	0.8995	0.5000	0.8221	0.5000	0.5000	0.0000

Table 2: Paired differences for two prespecified unadapted-model comparisons. Intervals use 2,000 source-stratified problem-group resamples. The weighted-macro A_{hazard} row covers seven of the eight sources. Pooled AUROC and global A_{hazard} use all eight. Paired differences therefore need not equal differences of the Table 1 point estimates.

Comparison	Metric	Difference	95% interval
Qwen–Gemma	Raw pooled AUROC	+0.0847	[0.0749, 0.0954]
	Global A_{hazard}	+0.0067	[-0.0025, 0.0159]
	Weighted-macro A_{hazard}	+0.0058	[-0.0059, 0.0186]
	Source-controlled IG	-0.0010	[-0.0075, 0.0053]
Muse–GLM	Raw pooled AUROC	+0.0026	[-0.0109, 0.0165]
	Global A_{hazard}	+0.0148	[0.0014, 0.0285]
	Weighted-macro A_{hazard}	+0.0597	[0.0391, 0.0794]
	Source-controlled IG	+0.0227	[0.0148, 0.0306]

5 Results

5.1 Absorbing labels amplify position leakage

We first score each step using only its relative position i/T_t; under the dataset-native (raw) labels pooled AUROC ranges from 0.4900 to 0.8805. We then impose absorption, holding each question, trace, score, trace length and first-error location fixed. Table 12 in Appendix J gives the per-corpus effect.

Already-absorbing sources do not change, the two sources with the highest recovery rates show the largest increases, and the source with low but nonzero recovery moves slightly the other way, so absorbing relabeling can create a large positional advantage without the direction being guaranteed for every corpus. Med-PRM supplies a paired annotation control: on the same 25,667 traces and 150,810 steps the relative-position baseline ranges from 0.4981 under its soft labels to 0.8281 under its Llama labels, with questions, steps and trace lengths identical and only the annotation changing.

5.2 Text-blind controls and unadapted-model rankings

Under pooled raw AUROC the absolute-index control scores 0.7534 and the source prior 0.8995, so counting steps outranks four of the five models and knowing only which corpus a trace came from outranks all five. Both receive exactly 0.5 from the source-aware A_{hazard} variants, and the source prior also receives 0 source-controlled IG. Global A_{hazard} does not suffice alone, since it still awards the source prior 0.8221. Macro aggregation, the usual repair, does not rescue the AUROC family either: under mean-macro absorbing AUROC the relative-position control scores 0.7981 and ranks first of the eight entries, above every model. It stays above chance under the first-error metrics too, because trace length varies within a source–index stratum, which Section 5.3 removes. Appendix G has the full raw and absorbing table with the control rows included.

For Qwen and Gemma, the large pooled-AUROC gap disappears under metrics that control first-error position and source. For Muse and GLM, pooled raw AUROC is inconclusive while every

Table 3: Restrictions applied cumulatively to pooled AUROC on the held-out set. Rows 5 and 6 are alternative aggregations of the same comparisons; row 7 adds length strata on top of row 5.

Estimand	Gemma	gpt-oss	Src. prior	Rel. pos.	Abs. index
Pooled AUROC, raw labels	0.9248	0.8559	0.8952	0.6126	0.6270
+ absorbing labels	0.9151	0.8528	0.9140	0.6208	0.6146
+ first-error risk set	0.8754	0.8066	0.8130	0.4964	0.3603
+ exact-index strata	0.8932	0.8152	0.8340	0.6758	0.5000
+ source strata, weighted	0.7925	0.7324	0.5000	0.6650	0.5000
+ source strata, mean	0.7963	0.7269	0.5000	0.6803	0.5000
+ trace-length strata	0.7723	0.7304	0.5000	0.5000	0.5000

reported first-error metric favors Muse. Neither pair is a sign reversal; in each the metric decides whether a difference is supported at all, which is already enough to change which model an evaluation would select. Section 5.4 exhibits the stronger case, where two estimands resolve the same comparison with opposite signs.

5.3 Ablating the estimand

Weighted-macro A_{hazard} differs from pooled AUROC in four independent ways: the label convention, the step population, index stratification, and source stratification. Table 3 imposes them one at a time on the held-out set, so each shortcut can be attributed to the single restriction that removes it.

Absorbing relabeling moves the first two rows by at most 0.0188, because only 1,558 held-out labels change, and it cannot move a later row at all: the risk set and the event indicator depend on a trace only through f_t, which the rewrite leaves fixed, so A_{hazard} is exactly invariant to the convention this paper is about, and the first-error index agrees under both conventions in all 7,280 traces of the held-out and validation sets combined. Restricting to the *first-error risk* set lowers every entry and drives the absolute-index control to 0.3603, because clean traces contribute late non-events that outrank the events of shorter erroneous traces. *Exact-index strata* return that control to exactly 0.5, and *source strata* drop the source prior from 0.8340 to 0.5000. Each control is eliminated by exactly one restriction and both verifiers survive all of them; Appendix L.1 has the full grid.

Trace-length strata are not an optional refinement. ActPRM and PRM800K stop annotating at the first error for every erroneous trace, so there the event is always the last retained step and relative position nearly identifies it. That control reaches A_{hazard} 0.9674 on ActPRM and 0.9649 on PRM800K, above Gemma at 0.8837 and 0.8458, gpt-oss at 0.7492 and 0.7634, and Qwen at 0.8514 and 0.7487. On the two largest sources in the evaluation set, reading no text beats every verifier we can measure, and the aggregate hides it. Adding length strata returns the control to exactly 0.5000 in every source while Gemma holds 0.8493 and 0.8147. The cost is efficiency, with comparable pairs falling from 1,300,848 to 194,780, so we report the length-stratified figure beside the primary one rather than in place of it.

The Mann–Whitney statistic sums over positive–negative pairs, so partitioning the pairs partitions it exactly. Writing S for pairs sharing a source and an index, D for pairs sharing a source only, and X for pairs crossing sources,

$$A_{\text{pooled}} = \frac{P_S A_S + P_D A_D + P_X A_X}{P_S + P_D + P_X}. \tag{7}$$

Under absorbing labels the held-out set has 371,765,485 pairs, of which S carries 0.59%, D 5.58%, and X 93.84%. The bucket holding almost all the mass is the one a text-blind scorer wins: the source prior reaches $A_X = 0.9412$ while receiving exactly 0.5 in both S and D, so its pooled absorbing AUROC is 0.9140 against Gemma's 0.9151, which is 99.7% of Gemma's above-chance discrimination, and it exceeds gpt-oss at 0.8528. Fitting the same prior on the validation split instead and applying it unchanged raises it to 0.9228, above Gemma, so the comparison does not depend on fitting the control to the test labels. Deleting any one source leaves the S share between 0.53% and 1.08%, and deleting any one of Math-Shepherd, MMLU-Pro-CoT or Qwen ProcessBench makes the prior outrank Gemma outright. The same accounting on the independent eight-source validation set puts 1,078,795 of the 1,189,004 comparisons that global A_{hazard} makes, or 90.7%, across sources,

Table 4: Held-out point estimates and 95% paired-bootstrap intervals on common score support. IG is in bits per risk-set step.

Model	Raw mean-macro AUROC	Weighted-macro A_{hazard}	Source-controlled macro IG
Qwen3.6-27B	0.8355 [0.8217, 0.8481]	**0.8385 [0.8207, 0.8549]**	**0.1426 [0.1248, 0.1619]**
Gemma 4 31B	0.8230 [0.8088, 0.8348]	0.7936 [0.7710, 0.8158]	0.1037 [0.0879, 0.1210]
gpt-oss-20b	0.7319 [0.7173, 0.7461]	0.7320 [0.7111, 0.7518]	0.0718 [0.0569, 0.0867]
Muse Glimmer 30B	**0.8438 [0.8317, 0.8552]**	0.8337 [0.8142, 0.8501]	0.1313 [0.1122, 0.1511]
GLM-4.7-Flash	0.7672 [0.7537, 0.7801]	0.7883 [0.7720, 0.8035]	0.0911 [0.0782, 0.1045]

so the pattern is a property of multi-source evaluation rather than of one split. Appendix L.2 has the full decomposition.

5.4 Matched training trajectories

The matched Gemma and gpt-oss runs use the same data, effective batch, LoRA configuration, and 2,300-step schedule, so comparing their first and last scored checkpoints separates conventional improvement from first-error improvement. Table 16 evaluates it under five ranking estimands and three readouts a practitioner would act on, the latter defined in Appendix L.5. One of the three is thresholded rather than rank based, so it is not a restatement of any AUROC.

For Gemma every ranking estimand agrees that the run improved, but the size of the improvement falls by a factor of four as nuisance is removed, from $+0.1074$ under pooled AUROC to $+0.1011$ under mean-macro AUROC, $+0.0646$ under global A_{hazard} and $+0.0267$ under weighted-macro A_{hazard}. The sign changes only when the target changes from ranking steps to naming one. Accuracy at naming the step that first broke falls by 0.0755 with interval $[-0.1060, -0.0443]$ while pooled AUROC rises by 0.1074 with interval $[0.0999, 0.1147]$, and the thresholded readout is flat at -0.0036. The highest-scoring step is the first error for 43.5% of Gemma's erroneous validation traces at step 100 and 35.9% at step 2,300, while the share landing after the first error rises from 42.3% to 51.0%. Absorbing supervision rewards flagging every post-error step, and the verifier learns that rather than where the failure began. Appendix L.3 rules out ties and saturation, and shows the result is not an artifact of choosing endpoints.

gpt-oss gains less under pooled AUROC, and its argmax localization is unresolved at -0.0109 while ProcessBench accuracy on erroneous traces resolves at $+0.0327$. No ranking estimand contradicts pooled AUROC here, including the source-aware ones; what contradicts it is the localization readout, and the source-aware variants matter because they show how much of the pooled gain was nuisance before that contradiction appears. The attenuation replicates in a third family: Qwen's mean-macro AUROC gain over its own run is 4.0 times its weighted-macro A_{hazard} gain, against 3.8 for Gemma and 2.6 for gpt-oss, so in all three most of the apparent gain lives in comparisons the first-error estimand declines to make.

5.5 Uniformly selected trained models

Every selected model covers more than 99.95% of the 40,475 held-out steps. Raw pooled AUROC exceeds raw mean-macro AUROC by 0.0963–0.1244 and raw weighted-macro AUROC by 0.0976–0.1119, while weighted- and mean-macro AUROC differ by at most 0.0173, so cross-source comparisons explain most of the pooled gap. Changing only 1,558 post-error labels also changes the ordering: Muse leads all three raw AUROC summaries, whereas Gemma leads absorbing pooled and mean-macro AUROC. Appendix H.1 has the complete table.

For paired inference, we restrict all models to their 40,402 common finite scores. Table 4 reports three complementary, source-aware estimands.

Muse's raw mean-macro lead over Qwen is 0.0083 with interval $[-0.0029, 0.0200]$. Qwen's weighted-macro A_{hazard} lead over Muse is 0.0049 with interval $[-0.0098, 0.0214]$, and its source-controlled macro-IG lead is 0.0112 bits with interval $[-0.0032, 0.0262]$. None of these three leading pairwise differences is resolved at the 95% level. In contrast, observation-weighted source-

controlled IG favors Qwen by 0.0104 bits with interval $[0.0024, 0.0181]$; this comparison depends on weighting observations rather than sources equally.

Equal-source A_{hazard} is especially sensitive to Double Delta, whose held-out risk set contains 303 events and only four non-events. In 141 of 10,000 bootstrap draws, that source supplies no comparable pair, and dropping it moves the mean-macro lead from Gemma to Qwen. Aggregation alone is therefore enough to move the winner: across the three first-error aggregations in Appendix H, global A_{hazard} is led by Muse at 0.9121, mean-macro by Gemma at 0.7963, and weighted-macro by Qwen at 0.8381, so one set of scores names three different best models, and the two source-aware aggregations among them disagree with each other. We report the declared seven-source estimate and the six-source sensitivity together, and we do not treat mean-macro A_{hazard} as decisive here.

5.6 PUBLIC-PRM COMPARISON

On the 3,973 steps of the three mathematics sources scored by every system, pooled AUROC and A_{hazard} disagree on three of seven entries, with Kendall's $\tau = 0.810$ between the two orderings. Math-Shepherd 7B moves from fifth under pooled AUROC, where it sits below the relative-position control at 0.6614 against 0.6633, to third under A_{hazard}, above both Qwen public PRMs. Appendix K gives the full ranking and the several reasons it is not a leaderboard.

6 DISCUSSION

Under absorbing labels, pooled AUROC measures whether invalid prefixes rank above valid prefixes; it does not isolate the step that caused failure. Global A_{hazard} removes absolute-index comparisons but can still compare sources, and mixed-source IG can likewise exploit source identity. Mean-macro and weighted-macro A_{hazard} and source-controlled IG are preferable when source is a nuisance.

Metric choice must follow the target. Raw labels measure local correctness and can credit recovery, absorbing labels measure prefix validity, and first-error metrics exclude recovery by design. Our evidence supports source-aware ranking metrics for measuring first-error discrimination, and a direct readout for measuring where the error is, rather than one universal ranking or selection rule.

7 CONCLUSION

Pooled step AUROC does not merely risk rewarding position and source identity. Held-out set used in our experiments is dominated by sequences that reward position and source identity, this is an accounting fact rather than an empirical tendency. Out of the 371,765,485 positive–negative pairs, 0.59% compare steps of one source at one index and 93.84% across sources. A scorer that reads no text, fit on a separate split and applied unchanged, outranks a fine-tuned 31B verifier, 0.9228 against 0.9151 respectively. Because a rank statistic is a sum over pairs, it costs one pass over set of labels, requires no model, and can be run before any verifier is trained. It can be used without the text or image (Kaushik & Lipton, 2018; Goyal et al., 2017) being made concrete. A_{hazard} is inherently resistant to position leakage. It includes one event per erroneous trace, counts post-error suffix once, and a random control scores exactly 0.5 on our metric. Additionally, in our ablation experiments, removing the source prior and relative-position prior, results in random control scoring exactly 0.5 under while both verifiers survive all the steps in our ablation. Over the matched run, all five ranking estimands rise with intervals excluding zero, from $+0.0708$ under pooled AUROC to $+0.0363$ under weighted-macro A_{hazard}, while the first-error detection on trace decreases by 0.0335 and accuracy on ProcessBench increases by 0.0273, both also excluding zero.

We show that using a single metric pooled-AUROC can gives misleading results. In our experiments, we establish empirically, that a text-blind baseline can often exceed a trained verifier. This motivated us to propose a new metric, A_{hazard}, which is resistant to position-leakage by construction. Our metric, shows relative improvement over text-blind baseline in ablation studies. We suggest readers to include orthogonally meaningful metrics in model comparisons along with text-blind baselines as evidenced by our results.

AI USE STATEMENT

We used a generative AI assistant in a limited capacity, for polishing prose and for proofreading the manuscript. It was not used to generate research ideas, to design the experiments, or to produce the results. The authors reviewed the methodology, claims and numerical results.

REFERENCES

Paul D. Allison. Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13:61–98, 1982. doi: 10.2307/270718.

Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975. doi: 10.1016/0022-2496(75)90001-2. The AUC = Mann-Whitney identity, which is what makes a pair-weighted statistic an AUROC. Chosen over Hanley and McNeil (1982, Radiology 143(1):29–36, doi 10.1148/radiology.143.1.7063747) because this is the exact identity paper and the paragraph is a provenance tour. Hanley and McNeil is 17,638 Crossref citations against 1,093 here – swap if a co-author prefers the recognizable anchor.

Matt Y. Cheung, Ashok Veeraraghavan, Hanjie Chen, and Guha Balakrishnan. Conformal certification of reasoning trace prefixes. *arXiv preprint arXiv:2605.30085*, 2026. URL https://arxiv.org/abs/2605.30085. Rice University. PREPRINT.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168. OpenAI.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x. Pages 187–202 are the paper, 202–220 the read-paper discussion; 187–220 is the conventional citation.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Keyu Duan, Zichen Liu, Xin Mao, Tianyu Pang, Changyu Chen, Qiguang Chen, Michael Qizhe Shieh, and Longxu Dou. Efficient process reward model training via active learning. *arXiv preprint arXiv:2504.10559*, 2025. URL https://arxiv.org/abs/2504.10559. Sea AI Lab. Source of the ActPRM corpus. PREPRINT.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

GLM Team, Aohan Zeng, et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670. Pages follow the IEEE record matching this DOI; CVF open-access paginates the same paper 6904–6913.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017.

Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005. doi: 10.1111/j.0006-341X.2005.030814.x.

Holly Janes and Margaret S. Pepe. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. *American Journal of Epidemiology*, 168(1): 89–97, 2008. doi: 10.1093/aje/kwn099.

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5010–5015. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1546.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050. OpenAI. Preprint: arXiv:2305.20050.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, Yunlong Feng, and Zhijiang Guo. AutoPSV: Automated process-supervised verifier. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.16802. Section 3.3 uses the "first error location" strategy and formalises the absorbing rule as $y_t = 0 \Rightarrow y_{t'} = 0$ for all $t' > t$.

Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. What are step-level reward models rewarding? counterintuitive findings from MCTS-boosted mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 24812–24820, 2025. doi: 10.1609/aaai.v39i23.34663. URL https://arxiv.org/abs/2412.15904.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491. The pair-weighting the metric is built on. Added because the closing sentence named Mann-Whitney while every other ingredient in that paragraph carried a citation.

Meta Superintelligence Lab. Muse glimmer 30b model card. Hugging Face model card, August 2026. URL https://huggingface.co/meta-models/Muse-Glimmer-30B.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. Hard2Verify: A step-level verification benchmark for open-ended frontier math. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22502–22517. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1031. Salesforce AI Research.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Yiren Shen and Juan J. Alonso. A multi-fidelity double-delta wing dataset and empirical scaling laws for gnn-based aerodynamic field surrogates. In *AIAA SCITECH 2026 Forum*. American Institute of Aeronautics and Astronautics, 2026. doi: 10.2514/6.2026-0686. URL https://doi.org/10.2514/6.2026-0686.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1230.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275. DeepMind.

Ph. van Elteren. On the combination of independent two-sample tests of Wilcoxon. *Bulletin of the International Statistical Institute*, 37:351–361, 1960.

Hanyin Wang, Chufan Gao, Qiping Xu, Bolun Liu, Guleid Hussein, Hariprasad Korsapati, Mohamad El Labban, Kingsley Iheasirim, Mohamed Hassan, Gokhan Anil, Brian Bartlett, and Jimeng Sun. Process-supervised reward models for verifying clinical note generation: A scalable approach guided by domain expertise. *arXiv preprint arXiv:2412.12583*, 2024a. URL https://arxiv.org/abs/2412.12583. Source of the PRM-Clinic corpus. PREPRINT.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.510.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024c.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 27723–27730, 2025. doi: 10.1609/aaai.v39i26.34987. URL https://arxiv.org/abs/2404.05692. Introduces ReasonEval. Table 1 reports step-level AUC beside macro-F1, with the rationale that AUC is "a more balanced evaluation of performance across different threshold settings".

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16554–16571. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.837.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.547. Qwen Team, Alibaba.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.50. Qwen Team, Alibaba.

A Exact model identifiers

Table 5 pins the revision of every released checkpoint used in the unadapted-model diagnostic.

Table 5: Pinned model identifiers used in the unadapted-model diagnostic.

Model	Revision
Qwen/Qwen3.6-27B	6a9e13bd6fc8f0983b9b99948120bc37f49c13e9
google/gemma-4-31B-it	842da3794eaa0b77d5f08bae87a17459d91ff475
zai-org/GLM-4.7-Flash	7dd20894a642a0aa287e9827cb1a1f7f91386b67
openai/gpt-oss-20b	6cee5e81ee83917806bbde320786a8fb61efebee
meta-models/Muse-Glimmer-30B	a4e59da52a7bc87ae7251dd5545c0dd437c44b68

B VERIFIER PROMPT

The system instruction is:

> You review one step of a solution. You are given the problem and the solution up to and including the step under review. Explain in a short paragraph whether that final step is sound, naming the specific error if there is one. Then state your verdict on its own final line as exactly "Verdict: VALID" or "Verdict: INVALID."

The user message contains the problem, optional answer choices and source material, the solution prefix through step i, and the instruction `Review step i.` Training and generated scoring use the pinned model chat template. Forced-verdict scoring uses the same problem and solution prefix but appends the tokenized `Verdict:` marker directly. Generated scoring locates that marker in token identifiers rather than re-encoding text, and decodes greedily in BF16 with at most 320 new tokens and 4,000 characters per step.

C INCREMENTAL-INFORMATION DETAILS

Context length is measured with `cl100k_base` from `tiktoken` 0.14.0 over the question and solution prefix through the current step. System instructions, chat-template tokens, verifier rationales, verdicts, future steps, and final trace length are excluded. The structural model contains an unpenalized intercept, categorical absolute-index effects, and a natural cubic spline of $\log(1 + L_{t,i})$. Four knots are placed at fitting-fold percentiles 5, 35, 65, and 95, with linear tails; non-intercept coefficients receive an L_2 penalty with $\lambda = 1$. We clip p_{invalid} to $[10^{-6}, 1 - 10^{-6}]$, transform and standardize its log odds using fitting-fold statistics, and set $q_\phi(s_{t,i}) = \gamma x_{t,i}$ with $\gamma \geq 0$ and no intercept. Five-fold cross-fitting is grouped by $(\text{source}, \text{question identity})$.

D TRAINING CONFIGURATIONS

Gemma and gpt-oss use LoRA rank 32, scaling parameter 64, dropout 0.05, learning rate 2×10^{-4}, one-cycle cosine decay with warmup fraction 0.03, zero weight decay, microbatch size 2, gradient accumulation 4, two distributed ranks, maximum sequence length 4,096, seed 1234, and 2,300 optimizer steps. Checkpoints are saved every 100 steps.

Qwen uses the same LoRA and learning-rate settings, microbatch size 8, gradient accumulation 8, an 8,192-token microbatch ceiling, maximum sequence length 4,096, seed 1234, and 600 optimizer steps. Checkpoints are saved every 50 steps.

GLM uses the same LoRA and learning-rate settings, microbatch size 16, gradient accumulation 8, an 8,192-token microbatch ceiling, maximum sequence length 4,096, 512-token loss chunks, seed 1234, and 1,493 optimizer steps. Checkpoints are saved every 100 steps.

Muse uses the same LoRA and learning-rate settings, microbatch size 2, gradient accumulation 4, maximum sequence length and padded microbatch budget 8,192, 256-token loss chunks, seed 1234, and 12,968 optimizer steps. Checkpoints are saved every 200 steps.

E CHECKPOINT-SELECTION DETAILS

Table 6: Uniform checkpoint selection using raw-label mean-macro AUROC.

Model	Candidates	Selected step	Validation AUROC
Qwen3.6-27B	12	350	0.8504
Gemma 4 31B	23	1,600	0.8216
gpt-oss-20b	23	2,200	0.7435
Muse Glimmer 30B	65	8,200	0.8395
GLM-4.7-Flash	12	800	0.8166

The GLM candidate set contains the 12 checkpoints scored through step 1,200; its final three saved checkpoints were unavailable for selection.

F MATCHED TRAINING TRAJECTORIES

Table 7 gives the endpoints behind the changes discussed in Section 5.4.

Table 7: Change from step 100 to step 2,300 in the matched runs, on the full 1,891-trace eight-source validation set rather than the 352-trace selection subset. Pooled AUROC is given under both label conventions, because the two conventions disagree on 623 labels in 205 traces of this set. A_{hazard} needs no such split, being invariant to the rewrite.

	Pooled AUROC, raw			Pooled AUROC, absorbing			Global A_{hazard}		
Model	First	Last	Δ	First	Last	Δ	First	Last	Δ
Gemma 4 31B	0.8271	0.9345	+0.1074	0.8479	0.9370	+0.0891	0.8296	0.8942	+0.0646
gpt-oss-20b	0.8227	0.8720	+0.0493	0.8440	0.8850	+0.0410	0.8184	0.8253	+0.0069

G FULL UNADAPTED-MODEL RESULTS

Table 8 is the complete raw and absorbing table promised in Section 5.2, with the three text-blind controls included beside the five models.

Table 8: Full forced-verdict results for the five unadapted models and the three text-blind controls, on the same 15,301 steps. IG is in bits per risk-set step. Bold marks the three columns in which a text-blind control, rather than a model, ranks first of the eight entries.

Model	Pool abs.	Pool raw	Mean-macro abs.	Mean-macro raw	Global A_{hazard}	Mean-macro A_{hazard}	Weighted-macro A_{hazard}	Mixed IG	Src.-ctrl. IG
Qwen3.6-27B	0.7885	0.7653	0.7690	0.7618	0.8770	0.8082	0.8504	0.1369	0.0964
Gemma 4 31B	0.7108	0.6806	0.7840	0.7768	0.8704	0.7987	0.8441	0.1109	0.0975
Muse Glimmer 30B	0.6887	0.6689	0.7725	0.7745	0.8214	0.7960	0.8057	0.0791	0.0601
GLM-4.7-Flash	0.6978	0.6663	0.7649	0.7072	0.8067	0.7191	0.7453	0.0652	0.0374
gpt-oss-20b	0.5301	0.5314	0.6268	0.5839	0.5993	0.6341	0.6162	0.0126	0.0091
Absolute index	0.7523	0.7534	0.7271	0.6644	0.5000	0.5000	0.5000	0.0000	-0.0001
Relative position	0.6350	0.6368	**0.7981**	0.7409	0.6187	0.6498	0.6621	0.0012	0.0127
Source prior	**0.9261**	**0.8995**	0.5000	0.5000	0.8221	0.5000	0.5000	0.0790	0.0000

H ADDITIONAL HELD-OUT METRICS

H.1 AUROC AGGREGATION

Table 9 reports all three AUROC aggregations under both label conventions for the five selected checkpoints.

Table 9: Held-out AUROC under raw and absorbing labels. Mean-macro gives each of seven sources equal weight; weighted-macro weights sources by within-source pair count.

Model	Coverage	Raw pool	Raw weighted	Raw mean	Abs. pool	Abs. weighted	Abs. mean
Qwen3.6-27B	0.9996	0.9316	0.8301	0.8353	0.9116	0.8192	0.7840
Gemma 4 31B	0.9997	0.9248	0.8195	0.8229	0.9151	0.8072	0.7958
gpt-oss-20b	0.9996	0.8559	0.7440	0.7316	0.8528	0.7372	0.6640
Muse Glimmer 30B	0.9997	0.9478	0.8467	0.8442	0.8933	0.8192	0.7096
GLM-4.7-Flash	0.9996	0.8822	0.7846	0.7673	0.8632	0.7953	0.7556

H.2 FIRST-ERROR AND CALIBRATED METRICS

The first table below gives the first-error estimands on each model's own score rows, and the second gives the frozen-calibrator results.

Table 10: Additional held-out point estimates on each model's available score rows. Mean-macro A_{hazard} includes all seven sources; the sensitivity excludes Double Delta. Source-controlled IG reports both observation-weighted and equal-source aggregation over six sources.

Model	Global A_{hazard}	Mean-macro A_{hazard}	Mean-macro A_{hazard}, sensitivity	Weighted-macro A_{hazard}	Src.-ctrl. IG, weighted	Src.-ctrl. IG, macro
Qwen3.6-27B	0.9076	0.7797	0.8419	0.8381	0.0765	0.1433
Gemma 4 31B	0.8932	0.7963	0.8348	0.7925	0.0570	0.1049
gpt-oss-20b	0.8152	0.7269	0.7466	0.7324	0.0392	0.0711
Muse Glimmer 30B	0.9121	0.7633	0.8238	0.8335	0.0660	0.1318
GLM-4.7-Flash	0.8562	0.7160	0.7854	0.7881	0.0515	0.0904

Table 11: Mean-macro calibrated metrics on six sources. Calibrators are fit on validation and frozen for held-out evaluation. Lower log loss is better.

	Local raw target		First-error target	
Model	Bal. acc.	Log loss	Bal. acc.	Log loss
Qwen3.6-27B	0.7078	0.4613	0.6645	0.3819
Gemma 4 31B	0.6621	0.4808	0.6200	0.4141
gpt-oss-20b	0.6219	0.5484	0.5872	0.4630
Muse Glimmer 30B	0.7028	0.4551	0.6124	0.4050
GLM-4.7-Flash	0.6660	0.5446	0.6156	0.4469

I Uncertainty and resampling design

Corpus-audit intervals resample complete traces. The two prespecified unadapted-model comparisons use 2,000 paired, source-stratified resamples of problem groups. For the five selected trained checkpoints we use 10,000 paired, source-stratified problem-group resamples on the 40,402 held-out rows carrying a finite score from every model, and the same resample is used for all models. The matched-run comparisons of Section 5.4 use 2,000 draws of the same design on the 1,891-trace validation set, and the endpoint-averaged variant in Appendix L.3 uses 400, which is fewer because each draw recomputes every statistic for ten checkpoints rather than two. Resampling whole problem groups keeps alternative solutions to one question together, and stratifying by source keeps the composition of every draw equal to the evaluation set.

Every verdict in this paper is read from an unadjusted 95% percentile interval, and we call a difference resolved only when that interval excludes zero. We apply no multiplicity correction and report none. Each metric admits ten model pairs, so a difference whose interval only just excludes zero should be read as nominal, and we avoid resting any claim on one. Where a resample leaves a source with no comparable event and non-event pair, the mean-macro estimators withhold the draw rather than average over a smaller source set, and we report how often that happens.

IG intervals resample the existing out-of-fold contributions without refitting the nuisance models, so they describe sampling variation in the evaluation set and not uncertainty in the fitted structural baseline. Calibrated-result intervals condition on calibrators fit on validation and frozen before held-out evaluation.

J Absorbing-label intervention by corpus

Table 12 gives the per-corpus effect of imposing absorption on the text-blind relative-position baseline, holding each question, trace, score, trace length and first-error location fixed.

Table 12: Effect of imposing absorbing labels on a text-blind relative-position baseline. Recovery is the percentage of traces whose labels return to correct after an error.

Corpus	Recovery	Raw	Absorbing	Change
ActPRM	0.00%	0.8481	0.8481	0.0000
Double Delta	25.74%	0.6644	0.9768	+0.3124
Math-Shepherd	1.27%	0.6372	0.6370	-0.0001
MMLU-Pro-CoT	0.00%	0.7730	0.7730	0.0000
PRM800K	0.00%	0.8805	0.8805	0.0000
PRM-Clinic	49.71%	0.4900	0.7148	+0.2248
Qwen ProcessBench	0.00%	0.7425	0.7425	0.0000

K PUBLIC-PRM COMPARISON

This comparison is a separate diagnostic and is computed on the eight-source validation set of Section 4, not on the held-out set used elsewhere in Section 5. It is restricted to the 3,973 steps of PRM800K, Math-Shepherd and Qwen ProcessBench scored by every system, and it compares the retained public PRMs with our Gemma and gpt-oss verifiers and two text-blind controls.

Table 13: Metric-dependent rankings on the common mathematics subset. Superscripts give rank among the seven retained entries in each column.

Type	Scorer	Pooled AUROC	A_{hazard}
Ours	Gemma 4 31B	$0.9139^{(1)}$	$0.9248^{(1)}$
Ours	gpt-oss-20b	$0.8375^{(2)}$	$0.8445^{(2)}$
Public PRM	Qwen2.5-Math-PRM-7B	$0.7018^{(3)}$	$0.6398^{(4)}$
Control	Relative position	$0.6633^{(4)}$	$0.6328^{(5)}$
Public PRM	Math-Shepherd 7B	$0.6614^{(5)}$	$0.7303^{(3)}$
Control	Absolute index	$0.5590^{(6)}$	$0.5000^{(6)}$
Public PRM	Qwen2.5-Math-7B-PRM800K	$0.4353^{(7)}$	$0.3952^{(7)}$

Three of seven entries change rank and Kendall's τ is 0.810. The table is not an unbiased leaderboard, for three reasons that pull in opposite directions. Our two models were fine-tuned on traces drawn from these mathematical sources. Two of the three public systems were also trained on corpora represented here, since Math-Shepherd 7B is trained on Math-Shepherd data and Qwen2.5-Math-7B-PRM800K on PRM800K, so the subset is not neutral ground for them either. Finally the A_{hazard} column is the global variant, which Section 5.2 shows still admits cross-source comparisons; with only three sources present the source-aware variant is the statistic our own argument prefers, and we did not recompute it for the public systems under their native readouts. The comparison should therefore be read as evidence that the ordering depends on the metric, and not as a measurement of which verifier is better.

L ESTIMAND ABLATION

The grid and the pair budget below use the 5,389-trace, 40,475-step held-out set; the criterion-validity subsection instead uses the 1,891-trace validation set, because only there do 46 checkpoints exist. The ablation covers Gemma and gpt-oss and not the other three trained verifiers, because the grid, the pair budget and the length-stratified variant all require per-step scores, and only those two runs retained them. This is a coverage limit rather than a selection: the two were fixed by which artifacts survived, not by their results, and the pair shares in Appendix L.2 are properties of the evaluation set up to coverage and so describe all five. Where a per-source figure was recoverable for a third model we report it, as in the Qwen values in Section 5.3.

`ablation_metric_design.py` recomputes every table here from the released scores. It ships a self-test covering the invariance claims of Section 3.4, the exactness of equation (7), and the recovery properties a first-error estimand should have: an absolute-index-only scorer receives exactly 0.5 under all three aggregations, a source-prior scorer exactly 0.5 under both source-aware ones, a

relative-position scorer exactly 0.5 once length strata are added, and a first-error oracle 1.0. Those four are exact identities rather than empirical observations, which is the sense in which the estimand recovers a known truth that pooled AUROC does not.

L.1 Complete estimand grid

Table 14 is the full factorial behind the cumulative path of Table 3, so any ordering of the four restrictions can be read off rather than inferred.

Table 14: Every combination of label convention, step population, index stratification and source aggregation. The first-error label is defined only on the risk set. The three estimands reported in the main text are pooled raw AUROC, mean-macro raw AUROC and weighted-macro A_{hazard}.

Label	Population	Index	Source	Gemma 4 31B	gpt-oss-20b	Source prior	Relative position	Absolute index
Raw	all steps	exact	mean	0.8142	0.7148	0.5000	0.7084	0.5000
Raw	all steps	exact	pooled	0.9085	0.8356	0.8878	0.5624	0.5000
Raw	all steps	exact	weighted	0.7715	0.7023	0.5000	0.6451	0.5000
Raw	all steps	none	mean	0.8229	0.7316	0.5000	0.7494	0.6520
Raw	all steps	none	pooled	0.9248	0.8559	0.8952	0.6126	0.6270
Raw	all steps	none	weighted	0.8195	0.7440	0.5000	0.7621	0.6936
Absorbing	all steps	exact	mean	0.7848	0.7099	0.5000	0.6967	0.5000
Absorbing	all steps	exact	pooled	0.8958	0.8336	0.9114	0.5809	0.5000
Absorbing	all steps	exact	weighted	0.7679	0.6947	0.5000	0.6447	0.5000
Absorbing	all steps	none	mean	0.7958	0.6640	0.5000	0.8303	0.7389
Absorbing	all steps	none	pooled	0.9151	0.8528	0.9140	0.6208	0.6146
Absorbing	all steps	none	weighted	0.8072	0.7372	0.5000	0.8052	0.7384
First error	risk set	exact	mean	0.7963	0.7269	0.5000	0.6803	0.5000
First error	risk set	exact	pooled	0.8932	0.8152	0.8340	0.6758	0.5000
First error	risk set	exact	weighted	0.7925	0.7324	0.5000	0.6650	0.5000
First error	risk set	none	mean	0.7950	0.7293	0.5000	0.6682	0.5697
First error	risk set	none	pooled	0.8754	0.8066	0.8130	0.4964	0.3603
First error	risk set	none	weighted	0.8350	0.7590	0.5000	0.7708	0.6209

L.2 Pair-budget decomposition

Table 15 gives equation (7) bucket by bucket under all three label and population choices.

Table 15: Exact decomposition of pooled AUROC by pair type, equation (7). Counts and shares are over the rows Gemma scored, which is 99.97% of the set, so they are properties of the evaluation set up to coverage rather than of any scorer. The three bucket values recombine to the pooled value in the last row of each block to within 10^{-9}.

Pair bucket	Share	Gemma 4 31B	gpt-oss-20b	Source prior	Relative position	Absolute index
Raw labels, all steps, 350,264,872 pairs						
same source, same index	0.64%	0.7715	0.7023	0.5000	0.6451	0.5000
same source, other index	6.27%	0.8244	0.7483	0.5000	0.7740	0.7134
different source	93.08%	0.9327	0.8643	0.9245	0.6015	0.6221
pooled	100%	0.9248	0.8559	0.8952	0.6126	0.6270
Absorbing labels, all steps, 371,765,485 pairs						
same source, same index	0.59%	0.7679	0.6947	0.5000	0.6447	0.5000
same source, other index	5.58%	0.8113	0.7417	0.5000	0.8221	0.7635
different source	93.84%	0.9222	0.8604	0.9412	0.6086	0.6064
pooled	100%	0.9151	0.8528	0.9140	0.6208	0.6146
First-error labels, risk set, 83,393,548 pairs						
same source, same index	1.56%	0.7925	0.7324	0.5000	0.6650	0.5000
same source, other index	10.96%	0.8411	0.7628	0.5000	0.7858	0.6381
different source	87.48%	0.8811	0.8134	0.8578	0.4571	0.3230
pooled	100%	0.8754	0.8066	0.8130	0.4964	0.3603

L.3 The matched-run divergence in detail

Neither ties nor saturation explains Gemma's localization loss. Over its erroneous validation traces the argmax is tied for 0.5% of traces at step 100 and 0.0% at step 2,300, and the maximum exceeds 0.999 for 86.5% at step 100 and 49.3% at step 2,300. The later checkpoint is both less saturated and less often tied, and still localizes worse.

Comparing two endpoints could manufacture the result, because step 100 is Gemma's single most favorable localization checkpoint at 0.4345 against roughly 0.355 over the rest of the run. Averaging each statistic over the first five and the last five scored checkpoints instead, with 400 resamples of the same design, pooled raw AUROC rises by 0.0708 with interval $[0.0664, 0.0750]$, weighted-macro A_{hazard} by 0.0363 with interval $[0.0212, 0.0508]$, the ProcessBench-style harmonic mean is flat at -0.0049 with interval $[-0.0208, 0.0115]$, and localization falls by 0.0335 with interval $[-0.0532, -0.0142]$. The flat harmonic mean is cancellation rather than absence of movement: erroneous-trace accuracy rises from 0.5364 to 0.5636, a gain of 0.0273 with interval $[0.0077, 0.0485]$, while clean-trace accuracy falls from 0.7393 to 0.6779. The two first-error readouts therefore disagree in sign on the same checkpoints, and we report both rather than the one that suits the argument. The divergence attenuates by roughly half and remains two resolved intervals of opposite sign, which is why the abstract and conclusion quote the averaged figure.

Qwen's 600-step run is a replication rather than a member of the matched pair. On the same eight-source validation set its mean-macro raw AUROC rises from 0.7904 at step 50 to 0.8290 at step 600, a gain of 0.0386, and its weighted-macro A_{hazard} from 0.8145 to 0.8242, a gain of 0.0096. Per-step scores for that run were not retained, so its localization cannot be computed. The full trajectory table for the matched pair is in Appendix F, and it is separate from final checkpoint selection.

L.4 Matched-run divergence by estimand

Table 16 is the full version of the ladder quoted in Section 5.4.

Table 16: Change from step 100 to step 2,300 in the matched runs on the eight-source validation set, with 95% intervals from 2,000 source-stratified problem-group resamples. The upper block is ordered by how much nuisance the estimand admits.

Statistic	Gemma 4 31B	gpt-oss-20b
Pooled AUROC, raw	$+0.1074$ [0.0999, 0.1147]	$+0.0493$ [0.0419, 0.0572]
Mean-macro AUROC, raw	$+0.1011$ [0.0884, 0.1136]	$+0.0391$ [0.0234, 0.0545]
Global A_{hazard}	$+0.0646$ [0.0510, 0.0790]	$+0.0069$ [-0.0084, 0.0226]
Weighted-macro A_{hazard}	$+0.0267$ [0.0072, 0.0466]	$+0.0149$ [-0.0105, 0.0408]
Mean-macro A_{hazard}	$+0.0204$ [-0.0031, 0.0418]	-0.0229 [-0.0813, 0.0195]
Trace selection AUROC	$+0.0673$ [0.0497, 0.0850]	$+0.0563$ [0.0348, 0.0792]
ProcessBench, harmonic mean	-0.0036 [-0.0242, 0.0162]	$+0.0242$ [0.0003, 0.0470]
ProcessBench, erroneous only	$+0.0136$ [-0.0124, 0.0399]	$+0.0327$ [0.0045, 0.0612]
First-error localization	-0.0755 [-0.1060, -0.0443]	-0.0109 [-0.0462, 0.0226]

L.5 Criterion validity across 46 checkpoints

Each estimand is computed for all 23 Gemma and 23 gpt-oss checkpoints of the matched runs on the 1,891-trace validation set, then rank correlated with three downstream readouts. Localization accuracy is the earliest argmax over the whole trace; the ProcessBench-style figure applies that benchmark's own readout to our scores at a fixed 0.5 threshold; trace selection scores a trace by its maximum step score. Adjacent checkpoints are correlated, so the effective sample is smaller than 46 and these coefficients are descriptive rather than tests.

Table 17: Spearman rank correlation between each estimand and each downstream readout over the 46 matched-run checkpoints, and within each family.

	ProcessBench-style			Localization			Trace selection		
Estimand	All	Gemma	gpt-oss	All	Gemma	gpt-oss	All	Gemma	gpt-oss
Pooled raw AUROC	+0.648	+0.039	+0.573	-0.629	-0.264	-0.313	+0.919	+0.907	+0.703
Pooled absorbing AUROC	+0.568	+0.026	+0.484	-0.625	-0.177	-0.316	+0.906	+0.875	+0.713
Mean-macro raw AUROC	+0.646	+0.005	+0.292	-0.496	-0.048	-0.088	+0.908	+0.891	+0.594
Weighted-macro raw AUROC	+0.518	-0.041	+0.307	-0.566	-0.239	-0.227	+0.854	+0.859	+0.508
Global A_{hazard}	+0.443	-0.027	+0.145	-0.334	+0.075	-0.030	+0.791	+0.864	+0.492
Mean-macro A_{hazard}	+0.624	+0.080	-0.052	-0.414	+0.009	+0.129	+0.814	+0.708	+0.271
Weighted-macro A_{hazard}	+0.681	+0.148	+0.060	-0.461	+0.058	+0.053	+0.883	+0.834	+0.397

The table should be read as three separate questions, and it does not say that one estimand dominates. Against the ProcessBench-style readout, which is the closest published analogue of the task A_{hazard} targets, weighted-macro A_{hazard} has the highest coefficient at $+0.681$ and pooled raw AUROC $+0.648$. Against trace selection the order reverses, pooled raw AUROC reaching $+0.919$ against $+0.883$. That reversal is what this paper predicts rather than a problem for it: trace selection asks whether a whole trace contains an error, which is close to the prefix-validity target pooled AUROC measures, while the ProcessBench-style readout asks which step broke. An evaluation should use the estimand matching its target, and neither number licenses using one statistic for both.

Every estimand is negatively correlated with localization accuracy, pooled raw AUROC most strongly at -0.629 and global A_{hazard} least at -0.334. We do not read this as evidence that A_{hazard} predicts localization, because it does not. Across these 46 checkpoints, higher scores on every rank statistic we compute go with worse argmax localization, which is the same phenomenon Section 5.4 isolates on the matched run. The honest conclusion is narrower than a criterion-validity claim: A_{hazard} removes two shortcuts that pooled AUROC pays for, and it is less anti-correlated with first-error localization than pooled AUROC, but no threshold-free step-ranking statistic we tested is a substitute for measuring localization directly.