ANSWER MATCHERS DETERMINE WHAT TEST-TIME SAMPLING APPEARS TO DO

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling is often framed as a policy problem. The system chooses how many samples to draw, when to stop, and whether to rerank, assuming that the measured outcome reflects the policy. On open-ended factual tasks, however, the measured outcome often reflects the answer matcher instead. We use the exact decomposition $R_S = C_S + A_S$ to split plurality error into non-coverage and aggregation failure. A matcher false negative enters both terms with opposite S-dependence, creating an apparent conflict between candidate acquisition and selection. On Llama-3.1-8B/TriviaQA, one normalization step affects 46% of all generations. Correcting it moves single-sample accuracy $0.185 \to 0.645$, turns total error from an increase ($0.781 \to 0.869$) into a decrease ($0.354 \to 0.225$), and reverses the measured gain of adaptive stopping from -5.50 to $+1.53$ points. The distortion is predictable rather than idiosyncratic. For the exposed class of matching rules, it equals the fraction of a model's *correct* answers in the affected style times its accuracy. Varying style directly by prompt at fixed capability keeps the proportionality but gives a dataset-dependent coefficient between 0.80 and 1.02, so we state the law as proportionality rather than as unit slope. The loss separates contemporaneous 8B open-weight models by an order of magnitude. Plain exact match is exposed in 15 of 33 cells and SQuAD-family normalization in none. We release the canary that detects the failure, which standard parsing metrics miss. Two further cases show the same shape. A 7B process reward model does not separate from trace length, and a 4096-dimensional prompt representation does no separate from a single prefix likelihood.

1 INTRODUCTION

Increasing test-time computation can improve language-model reasoning. Self-consistency samples multiple reasoning paths and aggregates their final answers (Wang et al., 2023b); search, reranking, and revision also trade inference cost for quality (Snell et al., 2025). Input-adaptive methods allocate more samples to prompts predicted to benefit (Aggarwal et al., 2023; Li et al., 2024; Damani et al., 2025). Repeated-sampling work distinguishes oracle coverage from the performance of a selection rule. Coverage can keep rising while majority or reward-model selection plateaus (Brown et al., 2024), and imperfect verifiers impose a separate ceiling (Stroebl et al., 2026). The practical question is not only *whether* more inference helps, but also *when, where, and for how long* to spend it.

These claims depend on how outcomes are measured. On open-ended factual tasks, measurement runs through an answer matcher. A normalization routine and alias comparison decide whether a generated string counts as correct. That lexical matching understates open-domain QA accuracy unevenly across systems is established (Si et al., 2021; Bulian et al., 2022; Kamalloo et al., 2023; Wang et al., 2023a), and we take it as given rather than as our finding. What is not established is the form of the dependence. We show that the matcher, not the sampling policy, can determine the sign of the headline effect, and that the size of its distortion is proportional to the fraction of a model's *correct* answers in the affected style times its accuracy, with a coefficient near one (Figure 1).

We built a finite-budget audit of test-time sampling around the exact decomposition $R_S = C_S + A_S$, which splits plurality error at budget S into non-coverage (the correct class never appeared) and aggregation failure (it appeared and lost the vote). The decomposition makes a specific prediction about matcher error. A false negative removes a correct draw, which raises C_S, and turns the rejected

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.

string into a spurious competitor, which raises A_S. Because C_S decays geometrically in S while the inflated competitor mass does not, an exposed matcher produces exactly the curve that a genuine acquisition/selection conflict would.

We observe this on real generations. On the same saved Llama-3.1-8B outputs for TriviaQA we compare two matchers. The *exposed* matcher retains terminal punctuation after normalization and belongs to the class of lightly normalized exact-match rules; the *robust* matcher strips edge punctuation per token. They disagree on 46% of rows, all in the same direction. Under the exposed matcher, non-coverage falls from 0.781 to 0.238 between $S = 1$ and $S = 32$ while total error rises from 0.781 to 0.869, which reads as acquisition and selection moving in opposite directions, and adaptive stopping appears to lose 5.50 points. Under the robust matcher, single-sample accuracy is 0.645 rather than 0.185, total error falls from 0.336 to 0.225, stopping gains 1.53 points, and no canonicalizer we test reproduces the conflict.

The effect is not uniform across models. On the same dataset the two matchers differ by 0.45 in accuracy for Llama-3.1-8B and by less than 0.002 for Qwen3-8B and Gemma-3-4B. Matcher bias is a function of answer style, and specifically of the style of a model's correct answers, rather than of capability. Style tracks family and instruction tuning, so the bias looks like a quality difference between models, and it is invisible to the parsing diagnostics normally used to catch measurement problems. On the affected slice the unparsed rate is 0.0004 and the empty-extraction rate 0.0000.

Contributions. (1) We show that the sign of the measured acquisition/selection trade-off on open-ended factual tasks is determined by the answer matcher, and that matcher bias scales with answer style rather than model quality (Section 5, EXP 0). We give a one-line canary that detects the failure and that standard parsing metrics miss. (2) We use $R_S = C_S + A_S$ to localize where the bias lands in A_S, by inflating the mass of competing incorrect classes, and report finite-budget curves on six model$\times$task cells. Under the robust matcher no cell shows the opposite regime. (3) We find the same cheap-scalar pattern on the selection side. A 7B process reward model does not separate statistically from a free "prefer the shorter trace" heuristic in any of 20 cells, although it beats majority voting in 5 of them. (4) On the observation side, a one-dimensional prefix-NLL scalar outperforms every learned observer we train, including a 4096-dimensional frozen prompt representation, in 9 of 12 cells. A $+6.5$ AUROC-point trajectory gain that looks resolved under one source of resampling is 1.6 combined standard deviations once question and split resampling are both counted. Together, these results show that observability, verification, and measurement can each look stronger than they are. In each case, a cheap scalar explains most of the apparent effect.

2 RELATED WORK AND POSITIONING

Self-consistency aggregates sampled chains (Wang et al., 2023b); SelfCheckGPT and semantic-uncertainty methods use variation across generations to detect unreliability (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Kossen et al., 2024). Adaptive-Consistency and ESC stop as whole samples arrive (Aggarwal et al., 2023; Li et al., 2024); other test-time-scaling methods allocate compute by difficulty or reward, control reasoning length, or score reasoning with process supervision (Snell et al., 2025; Damani et al., 2025; Muennighoff et al., 2025; Lightman et al., 2024). Our C_S/A_S accounting does not claim that the coverage/selection distinction is new (Brown et al., 2024; Stroebl et al., 2026); its role here is to localize *where* a measurement artifact enters, which requires the two terms to be separable. Classical value-of-computation and anytime-valid inference provide the decision-theoretic context (Russell and Wefald, 1991; Howard et al., 2021; Ramdas et al., 2023; Quach et al., 2024).

Answer matching is a known problem, and we are not the first to say so. That lexical matching understates open-domain QA accuracy is established. Si et al. (2021) mine knowledge bases for alternative answers and gain several exact-match points. Bulian et al. (2022) formalise answer equivalence and train a learned judge. Kamalloo et al. (2023) re-evaluate twelve systems on NQ-Open by hand and find both a large absolute gap and, importantly, that the relative ranking of systems changes under human judgment. Wang et al. (2023a) release EVOUNA, human correctness labels for five systems on NQ and TriviaQA, and report the same reordering under human judgment. The existence and the approximate size of the effect are therefore prior work, and we cite them for it rather than rediscovering it.

Three things separate what follows. The first is a quantitative form. Prior work reports that the gap varies across systems, and we give the law it follows. The loss is proportional to the fraction of a system's *correct* answers in the affected style times its accuracy. The between-system fit has unit slope and replicates on EVOUNA's held-out human labels; style varied by prompt at fixed capability gives a dataset-dependent coefficient between 0.80 and 1.02. The second is the systems. Those studies evaluate retrieval pipelines and 2023-era API models, whereas the exposure differs by an order of magnitude between contemporaneous 8B open-weight generators of different families (Section 5). The third is the consequence, which propagates into test-time scaling through $R_S = C_S + A_S$. A false negative enters both terms with opposite S-dependence and so counterfeits an acquisition/selection conflict. No prior evaluation work touches that mechanism, because none of it studies repeated sampling.

Secondary theory, pruning corrections, and selective-inference results are deferred to Appendices A and F.

3 Setup

3.1 Trajectory law, filtration, and probes

Let $\mathcal{V}$ be the vocabulary, $x \in \mathcal{V}^m$ a fixed prompt, and θ fixed parameters. Given a prefix $y_{<t}$ the model deterministically returns logits $\ell_\theta(x, y_{<t})$; a decoding specification ψ induces a next-token law $q_{\theta,\psi}(v \mid x, y_{<t})$ (Holtzman et al., 2020). For run s, $\Gamma^{(s)} = (Y_1^{(s)}, \dots, Y_{T_s}^{(s)})$ with $Y_t^{(s)} \sim q_{\theta,\psi}(\cdot \mid x, Y_{<t}^{(s)})$, T_s the first EOS time or a maximum length. Independent seeds give $\Gamma^{(1)}, \dots, \Gamma^{(S)} \overset{\text{iid}}{\sim} \mathbb{P}^x_{\theta,\psi}$. Writing $P_t^{(s)}$ for the observed prefix of run s at time t, an online controller observes $\mathcal{F}_t = \sigma(P_u^{(s)} : u \le t,\ s \in [S])$ and must be $\mathcal{F}_t$-adapted.

Independent probes. Let A, B be conditionally independent root-sample pools and H_B a B-measurable event. Under the fixed-law model $I(H_B; \mathcal{F}_t^A \mid X, \theta, \psi) = 0$, so an independent-pool AUROC gain is not posterior refinement beyond the prompt. It is the out-of-fold increment of a *specified finite* predictor when Monte-Carlo features $\Phi(\mathcal{F}_t^A)$ are added; we call this computational probe value. Appendix M fixes that observer.

3.2 Answer measurement is part of the estimand

Let $a(\Gamma)$ extract a normalized answer and $c(a)$ its class. A matcher is a pair $(\mathrm{nrm}, \mathrm{cmp})$, a normalizer applied to both the generation and the gold aliases, and a comparison rule. Correctness is $\mathrm{cmp}(\mathrm{nrm}(a(\Gamma)), \{\mathrm{nrm}(\alpha)\})$, and the class of an incorrect answer is its normalized surface form. Both C_S and A_S below are therefore functionals of the matcher, not of the model alone. A false negative acts twice. It removes mass from the correct class *and* adds a spurious competitor.

We treat this as a first-class part of the estimand rather than a preprocessing detail, because Section 5 shows a single normalization step flipping the sign of the headline result. Section 5 also reports the canary we now use. It is the fraction of rows scored incorrect whose extracted answer differs from some gold alias only by punctuation or casing. Unlike unparsed and empty-extraction rates, it is sensitive to matcher false negatives by construction.

With ground truth $A^\star$, aggregation rule $\widehat{A}_T$, and empirical answer measure $\widehat{\nu}_T^{(S)}$, the two terminal events are $E = \mathbf{1}\{\widehat{A}_T \neq A^\star\}$ and $G_\gamma = \mathbf{1}\{U(\widehat{\nu}_T^{(S)}) > \gamma\}$, with U the normalized entropy over observed classes and $\gamma = 0.35$. Disagreement is label-free; correctness is not, and only correctness passes through the matcher. Keeping the two separate is the central methodological commitment of this paper.

4 Finite-Budget Accounting

Let q_x be the law of a canonical answer class from one root sample, $a^\star$ the truth, $p_x = q_x(a^\star)$, and $m_x = \max_{a \neq a^\star} q_x(a)$. For an S-sample plurality decision let $N_S^\star$ be the correct-class count and

define

$$C_S(x) = \mathbb{P}(N_S^\star = 0 \mid x), \qquad A_S(x) = \mathbb{P}(\widehat{A}_S \neq A^\star,\, N_S^\star > 0 \mid x). \tag{1}$$

Then $R_S(x) = C_S(x) + A_S(x)$ exactly and, under exact canonicalization, $C_S(x) = (1-p_x)^S$. Thus $N_S^\star = 0$ establishes non-coverage *at budget* S, not $p_x = 0$. Additional i.i.d. samples can improve acquisition whenever $p_x > 0$, whereas reweighting a fixed candidate set can affect only selection.

The decomposition is what makes matcher bias localizable. A false negative deletes correct draws, which raises C_S at every S; because the rejected string is then classed by its surface form, it also raises m_x, which raises A_S. Since C_S decays geometrically while the inflated competitor mass does not, the two terms move in opposite directions as S grows, and the observed curve acquires exactly the shape of a genuine acquisition/selection conflict. Section 5 shows this curve on real generations.

Asymptotics constrain the endpoints but not the finite-S path. If the correct class is the unique mode ($p_x > m_x$), plurality error converges to zero. If some incorrect class exceeds p_x, it converges to one. The boundary $p_x = m_x$ is governed by the tie rule (Corollary 3, restated with the pairwise large-deviation rate in Appendix A). None of these statements implies monotonicity at every finite S, and Section 5 reports a systematic finite-S violation at $S = 2$.

Within the same law, longer prefixes refine a posterior over a fixed terminal event, and the resulting bounded martingale supplies an exact expected-accounting identity (Theorem 1, Appendix A). We use it only to organize posterior information and to check predictor calibration. It compares no actions, certifies no stopping rule, and is not used anywhere in Section 5. The decision-theoretic formulation and the plug-in control-regret bound are also deferred (Appendix B).

5 Experiments

Exp 0 establishes the matcher effect and its style dependence. Exp 2 reports acquisition/selection curves under the robust matcher. Exp 1 asks whether sampled prefixes improve a strong finite observer. Exp 3 evaluates stopping and verification against matched fixed-S baselines. They are presented in that order, which follows the argument rather than the numbering.

Setup. All robustness cells use 256 questions. Independent-pool prediction uses 64 root samples split into disjoint 32/32 A/B pools. The strongest sample-free observer includes a frozen generator prompt representation in addition to text/metadata features; trajectory variants add likelihood/completion and then semantic summaries. Learned comparisons use grouped CV and paired question bootstrap. Plurality quantities use the expected accuracy of uniform random tie-breaking rather than a lexical tie rule. Model revisions, decoding, canonicalization, split construction, and sensitivity protocols are in Appendix M. Cap-hit rates differ across cells and are reported with every claim; the length-controlled reruns are in Appendix E.

Exp 0 How much of the measured effect is the answer matcher?

What we look at. Whether an apparent acquisition/selection conflict is a property of the model or of the matcher, and whether matcher bias is uniform across systems.

What we find. The exposed matcher retains the characters `. + - / %` after normalization, so a sentence-final period survives and breaks alias matching. The robust matcher strips edge punctuation per token and requires word-boundary rather than character-level suffix matching. Both are applied to the same saved generations with no regeneration. Across all factual pools, 17,128 rows are scored incorrect by the exposed matcher and correct by the robust one, and 400 character-level false positives of the exposed matcher are removed. *No* row scored correct by the exposed matcher is scored incorrect by the robust one, so the difference is one-directional, and we verify this in every pool.

The effect is concentrated. On Llama-3.1-8B/TriviaQA, 72% of extracted answers carry edge punctuation and 46% of all generations are false negatives of the exposed matcher; single-sample accuracy is 0.185 under it and 0.645 under the robust matcher. On Qwen3-8B and Gemma-3-4B, same dataset and same pair of matchers, the difference is below 0.002 (Figure 7a, Appendix E). The bias is therefore not measurement noise distributed across systems but a term correlated with answer style, and style tracks family and instruction tuning rather than capability.

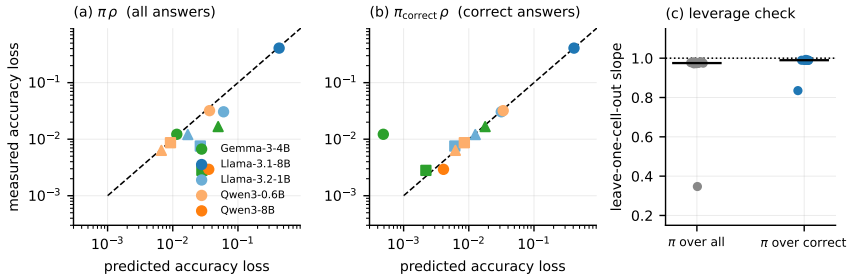


Figure 1: Matcher bias is predicted by the style of a model's *correct* answers. Measured accuracy loss against (a) style rate times accuracy over all answers and (b) over correct answers only, on 11 model-task cells with log axes; the dashed line is the identity. The rule pair shown is plain exact match against SQuAD normalization, neither of which we wrote. Panel (c) gives leave-one-cell-out slopes. The unconditional fit rests on a single high-leverage cell and collapses to 0.35 without it, while the conditional fit stays in $[0.84, 0.99]$. Style and correctness are strongly dependent for Gemma-3-4B and Qwen3-8B and nearly independent for Llama-3.1-8B, which is why the conditional form is the correct one.

Standard diagnostics do not detect it. On the affected slice the unparsed rate is 0.0004 and the empty-extraction rate 0.0000, so a standard parsing audit reports no problem. The canary of Section 3.2 fires at 0.46 under the exposed matcher (Figure 7b) and is 0.0000 on every factual slice under the robust one. We recommend reporting it alongside cap-hit and unparsed rates.

Which rule families are exposed. We apply six documented rules to the same generations over 33 pools spanning five models and three datasets. These are reimplementations of published descriptions and characterize rule *families* rather than named software releases. Ranked by mean canary rate, plain exact match on a lower-cased string is worst at 0.047, with 15 of 33 cells above 1%. The exposed matcher of this section is next at 0.044, also 15 cells. Every SQuAD-family variant is clean at 0.0013 and 0 cells (Appendix E). We therefore make no claim about results obtained with SQuAD-style normalization, and we do claim that evaluation scripts calling exact match on a lightly normalized string are exposed more than the matcher studied above.

Bias scales with the style of correct answers. A defect of this class can only remove credit from generations that were already correct and already in the affected style, so under independence the loss would be $\pi\rho$ for style rate π and accuracy ρ. That prediction fits well, with $R^2 = 0.988$ on the public rule pair, and it is untrustworthy. One cell carries the entire range, and removing it drops the slope from 0.975 to 0.347 and R^2 to 0.35. Style and correctness are in fact strongly dependent, and unevenly so. The ratio $\mathbb{P}(\text{style} \mid \text{correct})/\mathbb{P}(\text{style})$ runs from 0.04 on Gemma-3-4B/TriviaQA to 0.96 on Llama-3.1-8B/TriviaQA. Conditioning on correctness gives slope 0.990, intercept 0.001, $R^2 = 0.999$, and a leave-one-out slope range of $[0.835, 0.994]$, against $R^2 = 0.53$ for accuracy alone and 0.08 for style alone (Figure 1). The same fit on the exposed/robust pair of this section gives slope 1.001 with leave-one-out range $[0.886, 1.002]$.

The refinement matters for how the result should be read. Models answering in full sentences are not simply penalized. Qwen3-8B does so on 12% of answers and loses nothing, because only 1.5% of its *correct* answers carry edge punctuation. A model is penalized in proportion to how often it is right *and* in the affected style, a quantity that varies by an order of magnitude across these models and has no reason to track quality.

Moving style directly, at fixed capability. The fits above compare systems, so style and capability could be confounded, and the predictor is clustered in a way that makes a slope of one easy to obtain. We therefore vary answer style by prompt on one model, holding capability fixed by construction. Six instructions move π_{correct} from 0.16 to 0.67 on Llama-3.1-8B, and the predictor becomes continuous, with a largest gap of 0.11 against 0.60 on our pools and 0.45 on EVOUNA (Appendix E). Under that design the proportionality holds but the coefficient is not one and differs by dataset: 1.02 within TriviaQA ($R^2 = 0.94$), 0.80 within NQ-Open ($R^2 = 0.98$), 0.69 pooled ($R^2 = 0.75$). On TriviaQA

Table 1: The same generations under two matchers. Every row is computed from identical saved generations with no regeneration; only the answer labels differ. Intervals are 95% question-bootstrap.

Quantity (Llama-3.1-8B/TriviaQA unless noted)	**Exposed matcher**	**Robust matcher**
Single-sample accuracy	0.185	0.645
Non-coverage $C_1 \to C_{32}$	$0.781 \to 0.238$	$0.336 \to 0.125$
Aggregation failure A_{32}	0.631	0.100
Total error $R_1 \to R_{32}$	$0.781 \to 0.869$	$0.336 \to 0.225$
Opposite regime, any of 5 canonicalizers	yes	no
Stopping vs. matched fixed-S (pp)	$-5.50\ [-8.70, -3.25]$	$+1.53\ [+0.27, +2.72]$
Disagreement-target prevalence at γ=0.35	0.949	0.523
Matcher false-negative canary	0.4597	0.0000

the measured loss is 0.56 of the predicted, because gold aliases there are often multi-word and an answer with extra words still matches after normalization, so the predicate overstates π; on NQ-Open, where gold answers are short, the ratio is 1.15.

We therefore state the law as proportionality with a dataset-dependent coefficient rather than as unit slope. The between-system fits above are the case in which the predictor range is wide enough to hide that coefficient. What survives is the form, that the loss is governed by the product of style rate among correct answers and accuracy rather than by either alone, now supported by variation induced at fixed capability rather than only by comparison across systems. The sweep covers one model and two datasets, and whether the coefficient is a property of the dataset or of the cell is open.

Replication on labels we did not produce. EVOUNA (Wang et al., 2023a) provides human correctness judgments for five Open-QA systems on NQ and TriviaQA. Applying SQuAD-official normalization to those answers and conditioning on the *human* labels, which removes the proxy rule and the circularity with it, gives slope 1.017, intercept -0.019, $R^2 = 0.992$, and leave-one-out $[0.971, 1.083]$, against $R^2 = 0.21$ for accuracy alone (Appendix L). Two caveats belong with that number. The predictor support is narrow, since seven of ten cells have style rates above 0.9, where $\pi_{\text{correct}}\rho$ approaches ρ arithmetically, so this is a consistency check on held-out human labels rather than an independent test. And the defect SQuAD normalization is exposed to is verbosity rather than punctuation, so the two rules studied here are distinct defect classes reached by distinct predicates.

That distinction is testable, and it separates the law from a fit to any monotone covariate. Substituting the edge-punctuation predicate, which does not describe the defect SQuAD normalization is exposed to, degrades the fit to slope 0.848 with intercept $+0.123$ and $R^2 = 0.896$. The law holds with the predicate matching the defect class and fails with a mismatched one.

A sensitivity analysis over canonicalizers detects this failure only if the canonicalizers differ in their correctness test. Strategies that vary only how *incorrect* answers are grouped share one normalizer and one correctness test, produce identical curves, and give the appearance of robustness without testing it.

EXP 2 HOW DO ADDITIONAL SAMPLES TRADE ACQUISITION AGAINST SELECTION?

What we look at. Fair-tie curves C_S, A_S, R_S for $S \in \{1, 2, 4, 8, 16, 32\}$ on six cells, computed under the robust matcher.

What we find. R_S falls in all six cells (Figure 2). On Llama/TriviaQA, C_S falls $0.336 \to 0.125$ and A_S rises only $0 \to 0.100$, so total error falls $0.336 \to 0.225$, against $A_{32} = 0.631$ under the exposed matcher. The qualitative conflict flag is false for all five canonicalization strategies on both TriviaQA cells, including the permissive substring evaluator, and no cell of the six shows acquisition and selection moving in opposite directions.

Two observations remain. First, non-coverage remains the dominant term everywhere at small S and remains substantial at $S = 32$ on the open-ended cells ($C_{32} = 0.125$ on Llama/TriviaQA, 0.375 on Qwen/TriviaQA), so the finite-budget distinction between non-coverage and structural $p_x = 0$ continues to do work. Second, $R_2 > R_1$ on all three Llama cells ($0.164 \to 0.190$, $0.598 \to 0.604$,

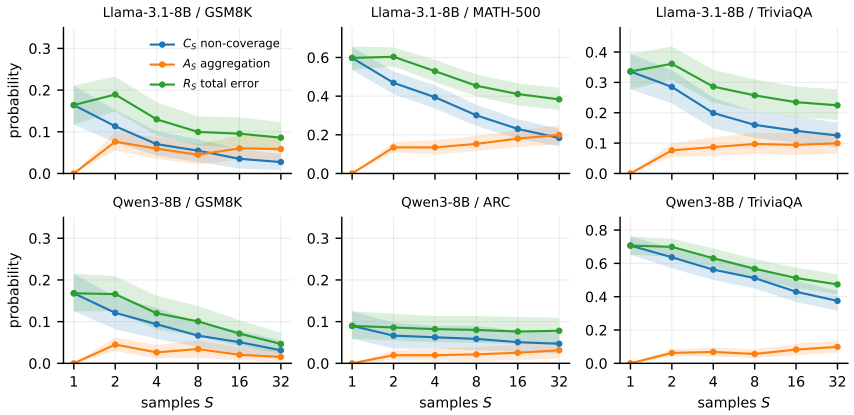


Figure 2: Fair-tie high-S audits under the robust matcher, 100 reveal orders and 2,000 question-bootstrap replicates per cell. $R_S = C_S + A_S$ holds exactly. R_S falls in all six cells; the opposite regime that the exposed matcher produces on Llama/TriviaQA does not appear. Note $R_2 > R_1$ on all three Llama cells, a finite-S non-monotonicity made visible by symmetric tie handling. Qwen/TriviaQA has a 63.8% cap-hit rate and is shown for completeness only.

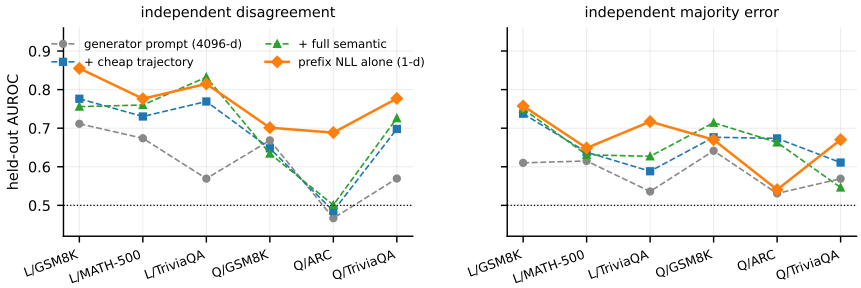


Figure 3: A one-dimensional prefix-NLL scalar against learned observers, six cells, 128-token checkpoint, grouped cross-validation by question. The scalar is contained in the "+ cheap trajectory" feature set, yet outperforms every learned model in 9 of 12 cells. On Qwen3-8B/ARC disagreement all learned variants sit at chance (0.47 - 0.50) while the scalar reaches 0.689.

$0.336 \to 0.361$) and not on the Qwen cells. The mechanism is the tie rule. At $S = 2$, a disagreement is a tie worth $1/2$ in expectation, whereas a single correct answer at $S = 1$ is worth 1. This is a concrete finite-S violation of the monotonicity that Corollary 3 explicitly does not assert, and it is visible only under symmetric tie handling; a lexical tie rule hides it.

EXP 1 DO SAMPLED PREFIXES IMPROVE A STRONG FINITE OBSERVER?

What we look at. Whether A-pool prefixes improve B-only prediction beyond a generator-aware, sample-free observer, and how that increment compares to the cheapest available trajectory feature.

What we find. Against the generator-prompt observer the trajectory features look useful. On Llama/GSM8K, the generator-prompt observer reaches disagreement AUROC 0.711 and adding cheap trajectory features reaches 0.776 ($\Delta = +0.065$, paired 95% CI $[+0.018, +0.111]$), with majority error $0.610 \to 0.738$. But that comparison is anchored on the wrong baseline. Prefix NLL alone is a single scalar already included in the cheap feature set. It reaches 0.855, and across the six cells and two targets it beats the best learned observer in 9 of 12 (Figure 3). On Qwen3-8B/ARC disagreement every learned variant sits at chance while the scalar reaches 0.689, which is a diagnosis of overfitting a 4096-dimensional representation on 256 questions rather than a property of the task.

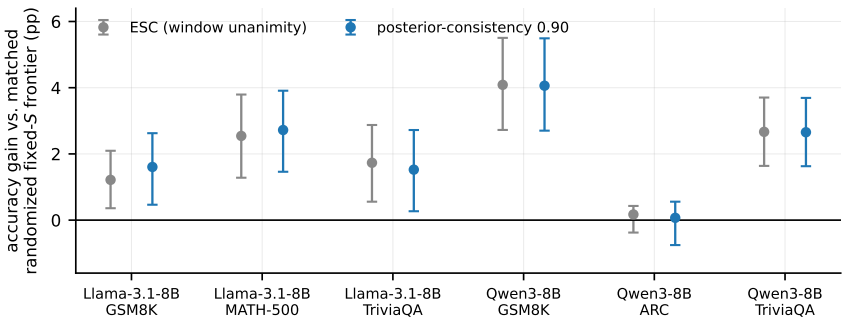


Figure 4: Posterior-consistency stopping and ESC against the randomized upper fixed-S frontier at policy-inclusive cost; 100 reveal orders, 2,000 question-bootstrap replicates, frontier recomputed inside each replicate. Both rules gain on five of six cells. The two are statistically indistinguishable everywhere, while ESC costs 0.004 - 0.011 ms of policy CPU against 0.03 - 0.40 ms for the posterior rule. Qwen/GSM8K, Qwen/ARC, Qwen/TriviaQA and Llama/MATH-500 use the original pools and remain length-sensitive.

The trajectory signal is still not an artifact of differential completion. On the 123 Llama/GSM8K questions where no A-pool path has completed by the checkpoint, prefix NLL still reaches AUROC 0.860 for disagreement and 0.798 for majority error.

Uncertainty must include both sources of resampling. A paired question bootstrap at one fixed A/B split and a count of positive repartitions (20 of 20) each hold one source fixed. Combining them, the split standard deviation (0.032) exceeds the question standard deviation (0.024), and the headline $+0.065$ is 1.62 combined standard deviations not resolved at 95%. Four of twelve cells exceed two combined standard deviations; the strongest is Llama/TriviaQA disagreement at 4.19. Under the exposed matcher that cell has prevalence 0.949 and would be excluded as degenerate; the imbalance is matcher-induced, the prevalence under the robust matcher is 0.523, and it is the one cell where the full semantic representation genuinely beats the cheap set ($+0.063$, $[+0.029, +0.099]$).

Exp 3 Which interventions move the measured quality-cost frontier?

What we look at. Whole-sample stopping and fixed-candidate verification, with symmetric tie handling and question-level uncertainty.

What we find. Under the robust matcher, stopping helps modestly and uniformly in direction. The pre-specified posterior-consistency rule gains $+1.61$ points on Llama/GSM8K ($[+0.47, +2.63]$) and $+1.53$ on Llama/TriviaQA ($[+0.27, +2.72]$), where the exposed matcher gives -5.50 ($[-8.70, -3.25]$); Qwen/ARC remains unresolved at $+0.07$ ($[-0.75, +0.56]$). The remaining cells use original pools with material cap-hit rates and are reported as length-sensitive.

The added complexity does not improve the result. ESC window unanimity is statistically indistinguishable from the posterior rule on all six cells and beats it on Llama/TriviaQA ($+1.73$ vs. $+1.53$), at roughly a fortieth of the policy CPU (Figure 4). Since ESC is an existing baseline (Li et al., 2024), the contribution of the Dirichlet posterior rule is not established.

Verification shows the same cheap-scalar pattern as observation. Against majority voting with uniform random tie handling, Qwen2.5-Math-PRM-7B gains up to $+5.94$ points (Llama/GSM8K, $S = 2$, $[+2.80, +9.44]$), and 5 of 20 cells are resolved positive. But weighting the identical candidate set by $1/(\text{tokens} + 1)$, a free scalar that requires no verifier, captures most of it. It gains $+3.15$ at the same cell, $+3.42$ against PRM's $+4.35$ on Qwen/GSM8K at $S = 2$, and within a tenth of a point of PRM at every S on that cell (Appendix E). Net of this control, *no* cell out of 20 has an interval excluding zero. A random per-candidate ranker gains approximately zero or less, so the control is not a degeneracy of majority voting at small S. We therefore withdraw the claim that fair-tie PRM scoring establishes a fixed-candidate selection gain. The data instead show that a 7B verifier is not separable from one scalar at these sample sizes.

The learned trajectory controller remains negative (mean -0.32 points over 404 configurations), and Stop-then-Verify does not robustly dominate its constituent baselines. Predictive observability is not by itself an action-value estimate.

6 LIMITATIONS

All results use the original candidate pools. Four cells have material cap-hit rates (Qwen/GSM8K 16.3%, Llama/MATH-500 13.2%, Qwen/ARC 5.4%, and Qwen/TriviaQA 63.8%), so they mix the matcher effect with truncation and are not used for headline claims. On Qwen/TriviaQA, accuracy is 0.819 on non-truncated generations versus 0.302 overall, and no 4096-token control exists yet. All robustness cells use 256 questions, so several reported differences correspond to a handful of questions and we give integer counts in Appendix E. The robust matcher is not itself validated by human annotation on our pools; we report it as a measurement whose failure mode we can detect, not as ground truth, and EVOUNA provides the only human reference we use. Finally, the ten-model factual-risk survey and the $10{,}000$-question decomposition (Appendices C and D) were computed with the exposed matcher and are reported with that caveat.

7 CONCLUSION

On open-ended factual tasks, what test-time sampling appears to do depends on the answer matcher. A single normalization step decides whether additional samples appear to help or hurt, and whether adaptive stopping appears to gain or lose, and the resulting bias falls on the model whose correct answers are full sentences rather than on the weaker model. The finite-budget decomposition $R_S = C_S + A_S$ makes this diagnosable. Matcher false negatives enter both terms with opposite S-dependence and create an apparent acquisition/selection conflict.

We see the same pattern on the other two axes. A 7B process reward model does not separate from trace length; a 4096-dimensional prompt representation does not separate from a single prefix likelihood. In all three cases, a cheap scalar explains most of an effect attributed to an expensive mechanism. Standard diagnostics do not flag these cases. The canary, the length control, and the scalar baseline are the instruments that do.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.761.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.20.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6qUUgw9bAZ.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.307.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927. ICML 2024 Workshop on Foundation Models in the Wild.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ndR8Ytrzhh.

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*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Stuart J. Russell and Eric Wefald. *Do the Right Thing: Studies in Limited Rationality*. MIT Press, 1991.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. What's in a name? answer equivalence for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9623–9629. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.757.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling, 2026. URL https://arxiv.org/abs/2411.17501. Earlier versions titled "Inference Scaling Flaws: The Limits of LLM Resampling with Imperfect Verifiers".

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-QA evaluation. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, 2023a. URL https://openreview.net/forum?id=UErNpveP6R.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

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*, pages 10495–10516. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.547.

Ethics Statement

The study analyzes reliability and inference efficiency; it does not provide a factuality guarantee. Low stochastic disagreement can coexist with systematic falsehood, and the learned trajectory controller is itself a negative result. High-stakes systems should therefore use independent verification, abstention, and domain-specific safeguards rather than treating agreement or early stopping as evidence of truth.

AI Use Statement

Language-model tools were used to code debugging, and to drafting manuscript text. All experiments were run by the authors on their own infrastructure. The authors are responsible for the claims, the experimental design decisions, and the final manuscript.

A Deferred Theory

This appendix collects the formal statements referenced in Sections 3.2 - 4. None of them is used to interpret the independent-pool probe gains, and none is an operational guarantee; they organize the two axes and are stated here so the main text can stay with the measurements.

A.1 Posterior resolution accounting on the token axis

Throughout this subsection $H \in \{0, 1\}$ is a *fixed* terminal event measurable at generation end, $R_t = \mathbb{P}(H = 1 \mid \mathcal{F}_t)$, and $V_t = R_t(1 - R_t)$.

A.2 Posterior resolution accounting

Theorem 1 (Posterior resolution identity)**.** *Let $t < s \le T$ and let $\{t_k\}_{k=0}^{K}$ with $t_0 = 0$, $t_K = T$ be any checkpoint grid.*

(i) $V_t - \mathbb{E}[V_s \mid \mathcal{F}_t] = \mathbb{E}[(R_s - R_t)^2 \mid \mathcal{F}_t]$.

(ii) Since H is binary, $V_T = 0$ almost surely, hence $\sum_{k=0}^{K-1} \mathbb{E}[(R_{t_{k+1}} - R_{t_k})^2] = \mathbb{E}[V_0]$. *For a fixed prompt x, R_0 is deterministic under the conditional law and this equals $R_0(1 - R_0)$.*

(iii) For any $\mathcal{F}_t$-adapted predictor $\widehat{R}_t$ with Brier risk $B_t = \mathbb{E}[(\widehat{R}_t - H)^2]$,

$$B_t - B_s = \mathbb{E}[(R_s - R_t)^2] + \mathbb{E}[(\widehat{R}_t - R_t)^2] - \mathbb{E}[(\widehat{R}_s - R_s)^2]. \tag{2}$$

The identity is the law of total variance specialized to a binary posterior and is valid for any filtration. Its value here is to place token-time observations on a common accounting scale. Part (ii) is an equality in expectation, not a pathwise conservation law. Part (iii) also shows why marginal calibration alone is insufficient. Identifying Brier improvement with posterior resolution requires zero or unchanged excess risk across checkpoints.

Scope for control. The identity concerns passive observation of a fixed H under a fixed continuation law. It does not compare actions that add root samples, branch, prune, change temperature, or change the prompt, because these actions induce different terminal decisions and laws. Nor does it by itself certify when the system should emit an answer. For the binary task of predicting H, any possible reduction in Bayes risk is already bounded by $\rho_t = \min(R_t, 1 - R_t)$, which is tighter than $\sqrt{V_t}$. We therefore use Theorem 1 only to organize posterior information and evaluate predictor calibration, not as an operational stopping guarantee.

A.3 The sample axis - large deviations and plurality limits

Let q_x be the law of a canonical answer class from one root sample, $a^\star$ the truth, $p_x = q_x(a^\star)$, $m_x = \max_{a \neq a^\star} q_x(a)$, and $\Delta_x = p_x - m_x$. The rate statement assumes a finite answer-class alphabet; the finite-budget decomposition of Eq. (1) does not require the multiclass rate. This matters for the open-ended factual cells of Section 5, where the observed class alphabet grows with S and Theorem 2 does not apply; those conclusions rest on the alphabet-free decomposition only.

Theorem 2 (Pairwise answer-order inversion)**.** *Let answer classes a, b have probabilities $u = q_x(a)$ and $v = q_x(b)$, with $u \neq v$, and let N_a, N_b be their counts in S i.i.d. draws. If $u > v$, then*

$$\mathbb{P}(N_b \ge N_a \mid x) = \exp\{-SI(u, v) + o(S)\}, \qquad I(u, v) = -\log\big(1 - (\sqrt{u} - \sqrt{v})^2\big). \tag{3}$$

If $v > u$, the same rate governs the rare reverse event $\mathbb{P}(N_a \ge N_b \mid x)$.

For a finite alphabet with a unique correct mode, a union bound over competitors yields plurality consistency. We do not claim a universal multiclass exponent when several competitors are active, nor an unrestricted result for an infinite open-ended answer alphabet. Smooth plug-in functionals $\Psi(\widehat{\nu}_S)$ form a separate regular class. Under the usual differentiability conditions their estimand is $\Psi(q_x)$ and their bias is typically $O(S^{-1})$; this is not an exhaustive taxonomy of hallucination definitions.

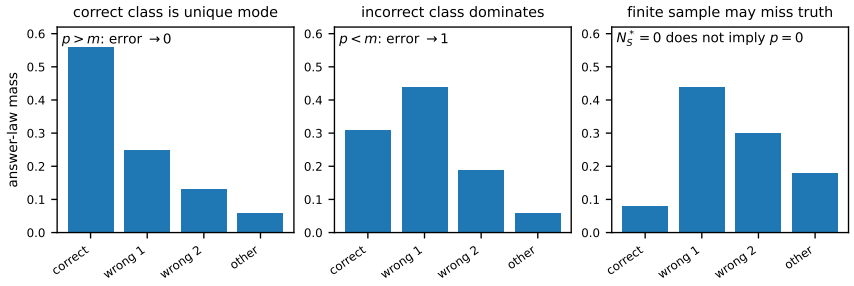


Figure 5: Three answer-law configurations. Panels (a) and (b) illustrate the asymptotic plurality limits when the correct class is respectively the unique mode or is dominated by an incorrect class; neither panel asserts monotonicity at every finite S. Panel (c) illustrates finite-budget non-coverage. Zero correct answers in an observed set implies $N_S^\star = 0$, not $p_x = 0$. The Hellinger expression governs a rare *pairwise* order inversion; multiclass plurality additionally depends on all competitors and tie-breaking.

Corollary 3 (Plurality limits and ties)**.** *If the correct class is the unique mode ($p_x > m_x$), plurality error converges to zero. If some incorrect class has probability strictly greater than p_x, plurality error converges to one and correct victory is exponentially rare. If $p_x = m_x$, the boundary has zero large-deviation rate and the limit depends on the number of tied modes and the specified tie-breaking rule. None of these asymptotic statements implies monotonicity at every finite S.*

Finite-budget decomposition (restated). Equation (1) of the main text gives $R_S = C_S + A_S$ exactly, with $C_S(x) = (1 - p_x)^S$ under exact canonicalization. For $S_2 > S_1$ with $C_{S_2} < C_{S_1}$, the descriptive ratio $\Gamma = (R_{S_1} - R_{S_2})/(C_{S_1} - C_{S_2})$ records how much recovered coverage converts into accuracy. Because both terms are matcher functionals (Section 3.2), Γ is only interpretable once the matcher canary is clean.

Corollary 4 (Mixing conditions and an exact margin decomposition)**.** *Let $q^{(1)}, \ldots, q^{(K)}$ be answer laws for one question under K conditions sharing the same truth, and $\bar{q} = K^{-1} \sum_k q^{(k)}$. Then*

$$\bar{\Delta} = \frac{1}{K} \sum_k \Delta^{(k)} + G, \qquad G = \frac{1}{K} \sum_k m^{(k)} - \max_{a \neq a^\star} \frac{1}{K} \sum_k q^{(k)}(a) \geq 0. \tag{4}$$

The mixture law has structural zero correct mass only if every condition has $p^{(k)} = 0$. At finite S, however, non-coverage remains possible even when all $p^{(k)} > 0$; for a stratified allocation $(n_1, \ldots, n_K)$ its probability is $\prod_k (1 - p^{(k)})^{n_k}$.

Equation (4) explains why diverse conditions can disperse competitors, while the finite-budget formula exposes the countervailing loss of within-condition votes.

A.4 What the framework classifies

Table 2 classifies methods by which stochastic object they change. The classification is descriptive; it does not substitute for matched-cost empirical comparisons.

The implication is narrower. Repeated sampling, reranking, and law-changing interventions are not interchangeable. Their relative value is an empirical action-value question, not a consequence of the martingale identity alone.

B A Decision-Theoretic View of Test-Time Interventions

The preceding quantities diagnose stochastic state; they do not by themselves define an adaptive policy. We use sequential decision theory only to state what a successful controller would have to estimate. A policy may continue paths, draw a root sample, branch, prune, change the generation

Table 2: Operations in the framework. Non-coverage is defined at a finite budget; only the statement $p_x = 0$ is structural.

Operation	What it can change
More i.i.d. root samples	Draws new candidates from the same q_x; can reduce C_S when $p_x > 0$ and can improve or worsen the finite plurality. It cannot alter q_x or repair structural $p_x = 0$.
ESC / vote-stability stopping	Stops acquiring whole samples when the observed vote is stable. It can save sampling cost, but agreement with a future vote is not a guarantee of correctness.
Confidence filtering / reranking	Changes selection on a fixed candidate set and therefore cannot reduce its non-coverage. Only A_S can move. Net of a free length-heuristic control on the identical candidate set we find no established gain (Appendix E).
Semantic entropy / SelfCheckGPT	Detection summaries of sampled variability; they do not allocate compute without an additional decision rule.
Longer generation / budget forcing	Changes the token-level continuation process and may change the induced answer law. The fixed-H identity is accounting within a specified law, not a performance ceiling. Our Qwen3-8B s1-style baseline does not uniformly improve the Pareto frontier (Appendix E).
Prompt diversification, retrieval, search	Changes the condition or candidate-generating law and can therefore affect both acquisition and selection.

condition, or stop. To put compute and terminal error in common units, define

$$J^\pi = \mathbb{E}\Big[\ell(\widehat{A}_{\tau_\pi}, A^\star) + \lambda C^\pi_{\text{total}}\Big], \tag{5}$$

where C^π_{total} is a declared compute metric (tokens, wall time, FLOPs, or a normalized combination) and λ is its exchange rate in units of terminal loss. Let $J^\star_t$ denote the optimal conditional cost-to-go from t. For a continuation action a, $Q_t(a) = \lambda c_t(a) + \mathbb{E}[J^\star_{t+1} \mid \mathcal{F}_t, a]$; under 0-1 loss the stop action has value $Q_t(\text{stop}) = \mathbb{P}(\widehat{A}_t \neq A^\star \mid \mathcal{F}_t)$ (plus any declared stop-side cost). Theorem 1 does not compare these counterfactual actions; action values must be estimated and calibrated directly. Proposition 8 in the appendix relates action-value error to control regret. The pruning correction in Appendix F applies only to path-factorized randomized pruning and does not justify bundle-coupled rules based on pairwise dispersion or cluster entropy. In particular, AUROC for disagreement is not an action value. Control depends on calibrated differences in expected terminal loss plus compute cost, and the stop action is correctness-relevant. The negative learned-controller result below is therefore a direct test of whether the observed trajectory signal is sufficient for control, not a contradiction of its predictive value.

C Factual Risk Across Ten Models

What we look at. Whether the factual-risk signal is an artifact of one model or one scale, evaluated at a fixed 128-token checkpoint after alias-aware relabeling.

What we find. After canonical relabeling, mean AUROC at 128 tokens spans 0.726-0.799 across ten models and three datasets, present at every scale and family, with most curves saturating by 128-256 tokens. Prevalence varies from 0.922 to 0.467, so AUROC must be read jointly with it. Pooled over the two 8B models, Brier improves from 0.177 to 0.140 by 128 tokens with ECE below 0.034 (Appendix Figure 11).

D Finite-Budget Decomposition on Closed-Book Factual Questions

What we look at. How the observed $S = 16$ candidate set and vote vary on open-ended factual recall.

What we find. Over 10,000 questions from five models, the correct class is unseen in 70.1% of the observed 16-sample sets, empirically modal in 24.6%, and observed but nonmodal in 5.4%. Majority error changes from 0.774 at $S = 1$ to 0.757 at $S = 16$. These figures establish severe finite-budget acquisition difficulty in the measured range; they do not identify structural $p_x = 0$ or prove that no larger budget can help.

Scale mainly reduces observed non-coverage on TriviaQA (from 0.52 to 0.36 between the reported 1B and 14B models), while the observed nonmodal share remains small. At $S = 1$, aggregation failure is zero by definition and rises to 0.077 by $S = 16$. The reported confabulation statistic

is approximately flat under its U-statistic estimator. This is one smooth functional rather than a taxonomy of all factual errors.

Observation cost. A semantic encoder evaluated at five checkpoints costs $4.22\times$ the estimated generation FLOPs for Qwen-0.6B and $1.96\times$ for Llama-1B, versus 0.15-$0.28\times$ for the 8-14B models (median 0.53). These FLOP ratios do not include latency, memory, or batching overhead. They motivate total-cost accounting rather than complete it.

E Robustness and Additional Controls

Strong prompt observer and repeated A/B partitions. We evaluate the finite-observer claim against several sample-free baselines. These include MiniLM prompt embeddings, generator-side prompt statistics, and a frozen generator prompt representation. On Llama/GSM8K, the generator representation reaches disagreement AUROC 0.711; adding cheap trajectory features reaches 0.776, with paired 95% CI $[+0.018, +0.111]$. Across 20 random 32/32 partitions of the same 64 samples, the cheap trajectory gain is positive in 20/20 splits (mean $+8.1$ AUROC points). The full semantic representation does not consistently beat the cheap trajectory set. For the disagreement threshold $\gamma \in \{0.20, 0.30, 0.35, 0.40, 0.50\}$, the Llama/GSM8K cheap-feature gain remains positive with positive paired intervals. Other tasks are more heterogeneous. Llama/TriviaQA disagreement is highly imbalanced at $\gamma = 0.35$ under the exposed matcher, with prevalence 0.949, which would exclude the cell. The imbalance is matcher-induced; under the robust matcher the prevalence is 0.523 and the cell is usable.

Truncation audit and length controls. The original 2048-token Qwen/GSM8K pool has a 16.3% cap-hit rate and the original Qwen/TriviaQA 256-token pool has 63.8% cap hits. A new 4096-token Qwen/GSM8K control reduces cap hits to 4.6% and raises single-sample accuracy to 93.8%; Qwen/ARC falls to 1.1% cap hits. Llama/MATH-500 remains length-sensitive even at 4096 tokens (12.6% cap hits), so its stopping and verifier results are secondary. Llama/GSM8K and Llama/TriviaQA have low original cap-hit rates (0.7% and 2.3%).

Stopping uncertainty, reveal-order robustness, and the ESC control. For every reported stopping cell, we use 100 reveal orders. Population uncertainty comes from 2,000 question-level bootstrap replicates. The fixed-S randomized upper frontier is recomputed inside each replicate, and reveal-order averaging is performed within the replicate. At policy-inclusive cost, the posterior-0.90 rule gives $+1.61$ on Llama/GSM8K ($[+0.47, +2.63]$), $+1.53$ on Llama/TriviaQA ($[+0.27, +2.72]$), $+2.72$ on Llama/MATH-500 ($[+1.46, +3.91]$), $+4.06$ on Qwen/GSM8K ($[+2.70, +5.49]$), $+2.66$ on Qwen/TriviaQA ($[+1.63, +3.69]$), and $+0.07$ on Qwen/ARC ($[-0.75, +0.56]$). The last four cells use the original pools and remain length-sensitive; the earlier length-controlled Qwen/GSM8K result was $+0.54$ ($[+0.04, +1.24]$), and the corresponding 4096-token reruns of the other three are the main outstanding item.

ESC window unanimity is the appropriate cheap baseline. Its gains are $+1.22$, $+1.73$, $+2.54$, $+4.09$, $+2.67$ and $+0.17$ on the same six cells, i.e. statistically indistinguishable from the posterior rule everywhere and larger on Llama/TriviaQA. Mean policy CPU is 0.004 - 0.011 ms for ESC against 0.030 - 0.396 ms for the posterior rule. We therefore do not claim that the Dirichlet posterior rule improves on ESC.

The two matchers and canonicalization sensitivity. The exposed matcher applies NFKC folding, lower-casing, article removal, and then retains the characters `. + - / %`. Two consequences follow. A sentence-final period survives normalization, so `"Anything Goes."` does not match the alias `"Anything Goes"` under either exact or suffix comparison. And article removal runs *before* punctuation handling, so the word-boundary pattern matches inside dotted abbreviations, mapping `"U.S.A."` to `"u.s. ."`. Separately, its suffix rule compares characters rather than word boundaries and accepts `"first practical telephone"` as a match for the alias `"phone"`.

The robust matcher reorders the two steps and strips edge punctuation per token while preserving internal punctuation (`3.5` is unchanged). It restricts article removal to whole tokens, and its suffix rule requires a word boundary. On the saved pools, 17,128 rows are scored incorrect by the exposed

matcher and correct by the robust one, 400 character-level false positives of the exposed matcher are removed, and no row moves in the other direction for punctuation reasons in any of 33 pools.

A sensitivity analysis over canonicalizers is only informative if they differ in their correctness test. Of the five strategies we report, four differ only in `wrong_mode`, which controls how *incorrect* answers are grouped into classes; built on the exposed normalizer they share one correctness test and produce four identical curves, which looks like robustness and is not. Under the robust normalizer, the qualitative conflict flag is false for all five strategies on both TriviaQA cells. Llama/TriviaQA gives $R_1 \to R_{32}$ of $0.383 \to 0.281$ (exact), $0.336 \to 0.225$ (suffix) and $0.254 \to 0.170$ (contains); Qwen/TriviaQA gives $0.715 \to 0.487$, $0.707 \to 0.474$ and $0.438 \to 0.234$ respectively.

Exposure of published matching rules. We reimplement six rules from their documented descriptions and apply them to the same stored generations over 33 pools covering five models and three datasets. The rules are the exposed matcher, the robust matcher, plain exact match after lower-casing and whitespace collapse, SQuAD-official normalization with exact comparison, the same with accent folding, and SQuAD normalization with containment. These are reimplementations, so they characterize rule families rather than any named release. A claim about a specific harness requires pinning and running it.

rule	mean canary	max canary	cells $> 1\%$ (of 33)
plain exact match	0.0466	0.4003	15
exposed matcher (Sec. 5)	0.0436	0.3995	15
SQuAD official	0.0013	0.0076	0
SQuAD with accent folding	0.0013	0.0076	0
SQuAD with containment	0.0001	0.0017	0
robust matcher (Sec. 5)	0.00004	0.0007	0

Stripping punctuation removes this defect class entirely, so SQuAD-family normalizers are not implicated. Plain exact match is exposed more than the exposed matcher of Section 5.

Style interaction and its leverage. For each factual pool we compute the style rate π, meaning the extracted answer carries edge punctuation after normalization, together with the reference accuracy ρ and the measured accuracy loss under the exposed rule. Under independence the loss should be $\pi\rho$. Fitting measured loss on that predictor over 11 cells gives slope 0.975 and $R^2 = 0.988$ for the public rule pair, but the fit is carried by one cell. Llama-3.1-8B/TriviaQA sits at $x = 0.43$ while the other ten lie below 0.06, and removing it drops the slope to 0.347 and R^2 to 0.349. Independence does not hold. The ratio $\mathbb{P}(\text{style} \mid \text{correct})/\mathbb{P}(\text{style})$ ranges from 0.042 on Gemma-3-4B/TriviaQA to 0.960 on Llama-3.1-8B/TriviaQA, so edge punctuation sits mostly on long incorrect answers for some models and is independent of correctness for others. Replacing π by π_{correct} gives slope 0.990, intercept 0.001, $R^2 = 0.999$, and a leave-one-cell-out slope range of $[0.835, 0.994]$. On the exposed/robust pair the corresponding numbers are 1.001 and $[0.886, 1.002]$. Accuracy alone gives $R^2 = 0.53$ and style alone $R^2 = 0.08$. We report the leave-one-out range in the main text because the unconditional fit is exactly the kind of result that looks strong and rests on one point.

The style predicate is a single function in the released code, so the same analysis applies to a different defect class by replacing it. We have not done so, and in particular we do not claim the interaction transfers to defect classes that SQuAD-family normalizers are exposed to, such as incomplete alias sets.

Matcher canary. The audit script now reports, per factual slice, the fraction of rows scored incorrect whose extracted answer differs from some gold alias only by punctuation or casing, and warns above 1%. On Llama/TriviaQA this was 0.4597 before the fix and 0.0000 after; the unparsed rate (0.0004) and empty-extraction rate (0.0000) were within tolerance throughout and the automated audit therefore reported no parser problem. We recommend the canary as a standard companion to cap-hit and unparsed rates for any open-ended factual evaluation.

Regime-label power. Labels based on the sign or an interval for Δ_x change with estimation-pool size, especially on open-ended factual tasks. Main claims therefore use directly estimable C_S, A_S, R_S curves rather than treating helpful/harmful/boundary labels as intrinsic prompt properties.

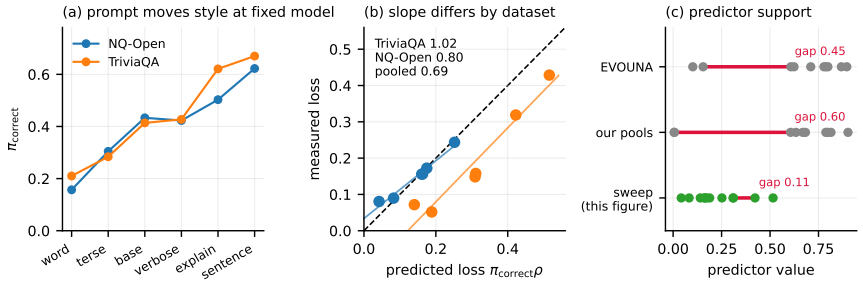


Figure 6: Style varied by prompt on Llama-3.1-8B, capability fixed by construction, twelve cells. (a) Six instructions move the style rate among correct answers by a factor of four. (b) Measured against predicted loss, with the within-dataset fits; the dashed line is the identity. The proportionality holds within each dataset with $R^2 \geq 0.94$, but the coefficient is 1.02 on TriviaQA and 0.80 on NQ-Open, and pooling gives 0.69. (c) The predictor is continuous here, with a largest gap of 0.11, where the between-system designs of Figure 1 leave gaps of 0.45 and 0.60. The weaker pooled slope is therefore not noise; it is what the clustered designs could not resolve.

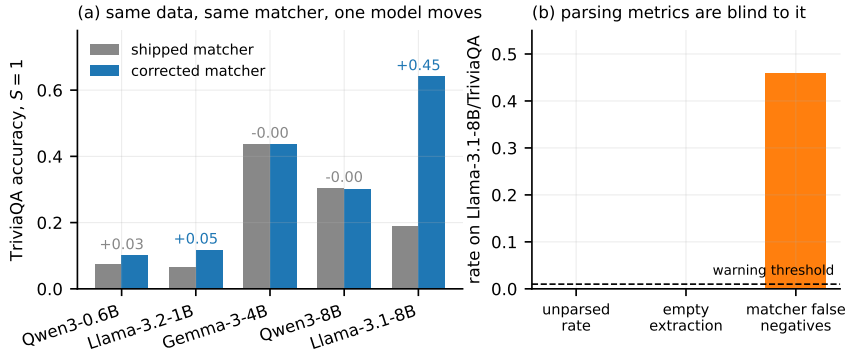


Figure 7: Answer-matcher bias is a function of answer style, not capability. (a) TriviaQA single-sample accuracy on identical generations under the exposed matcher (grey) and the robust matcher (blue). The two differ by 0.45 for Llama-3.1-8B and by less than 0.002 for Qwen3-8B and Gemma-3-4B; Llama-3.1-8B answers in full sentences, and the exposed matcher retains the sentence-final period. (b) Standard parsing diagnostics do not detect this. The answer parses cleanly but is rejected. The proposed canary, the fraction of *incorrect*-scored rows differing from a gold alias only by punctuation or casing, fires at 0.46.

Fair-tie process-reward selection with a free control ranker. We score fixed GSM8K and MATH-500 candidate sets with Qwen/Qwen2.5-Math-PRM-7B (Zhang et al., 2025). Majority accuracy uses uniform random tie-breaking, so $S = 2$ disagreements are not resolved lexicographically. Against that baseline the PRM gains up to $+5.94$ points (Llama/GSM8K, $S = 2$, $[+2.80, +9.44]$), and 5 of 20 cells with $S > 1$ are resolved positive.

Because the candidate set is fixed, any per-candidate scalar defines a comparable reranker. We therefore add two controls on the identical sets. One weights by $1/(\text{tokens} + 1)$ ("shorter trace"), and the other uses a fixed random score per candidate. The random ranker gains approximately zero or less at every cell, so the comparison is not a degeneracy of majority voting at small S. The length heuristic, however, tracks the PRM closely. On Qwen/GSM8K, it gives $+3.42/+1.86/+1.09/+0.62$ points at $S = 2/4/8/16$ against the PRM's $+4.35/+2.17/+1.40/-0.08$. Net of the length control, no cell out of 20 has a paired interval excluding zero. Non-coverage is identical before and after any reranking by construction. We therefore report $\Delta_{\text{PRM}-\text{length}}$ rather than $\Delta_{\text{PRM}-\text{majority}}$ as the quantity of interest, and we do not claim an established fixed-candidate selection gain.

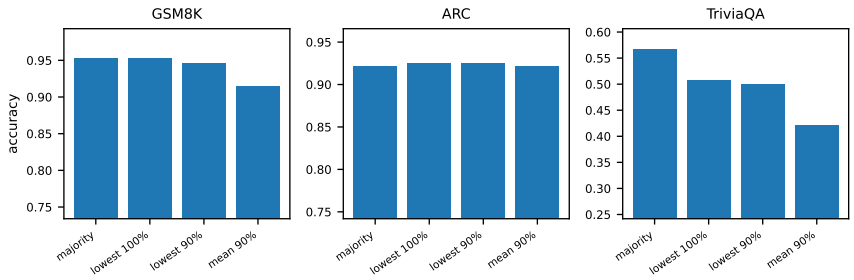


Figure 8: DeepConf-style token-confidence replay on the Qwen3-8B headline candidate pools (Fu et al., 2025). Reranking leaves finite-budget non-coverage unchanged by construction and does not consistently improve majority accuracy. This is not a faithful reproduction of the official DeepConf procedure.

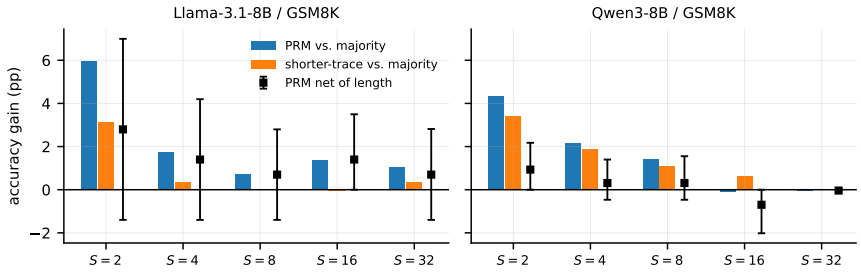


Figure 9: Process-reward reranking against a free control on the identical candidate set. Bars compare each ranker with fair-tie majority voting; squares show PRM net of the length heuristic with paired question-bootstrap intervals. Across all 20 cells ($S > 1$, four generator/task pairs) the net interval excludes zero in none.

Test-set sizes are 143 prompts (Llama/GSM8K) and 151 - 161 elsewhere, not 256; a one-point difference is roughly 1.5 questions.

Verifier serving caveats. Net of the length control the verifier experiment establishes no fixed-candidate selection gain, and it establishes no serving efficiency result either. Padded micro-batches of size 2, 4, and 8 did not numerically reproduce the unpadded batch-1 rewards, so final scores use batch 1. A separate online Qwen latency benchmark also failed to reproduce a favorable offline matched-cost comparison. We therefore avoid a deployment-level Pareto claim.

Naive Stop-then-Verify composition. A calibrated vote-margin rule invokes PRM only on ambiguous stopped prompts. Its best gain over an envelope containing fixed majority, fixed PRM, stopping-only, and always-PRM baselines is $+0.42$ points on Qwen/MATH-500, essentially zero on Qwen/GSM8K, and not stable on Llama. The negative result shows that the tested ambiguity threshold is not a reliable proxy for verifier action value; it does not establish that a richer action-value model would succeed.

s1-style budget-forcing control. Table 3 reports the Qwen3-8B reasoning-length baseline. The intervention suppresses one attempted end-of-thinking and resumes with "Wait"; because the original s1 uses a separately trained model, this is an s1-style intervention rather than a faithful reproduction. No setting uniformly improves accuracy while reducing tokens. The largest observed gain is MATH-500 at 2048 tokens ($+3.13$ points at $+4.5\%$ mean tokens), while GSM8K at 512 tokens is strongly harmful (-8.33 points at $+5.3\%$ tokens). Each row uses 96 prompts, so one point is approximately one prompt.

Table 3: Qwen3-8B s1-style wait forcing versus ordinary generation at the corresponding reasoning budget. ΔAcc is forced minus ordinary accuracy; ΔTok is the relative change in mean generated tokens.

Task	Budget	Acc. ordinary	Acc. +Wait	ΔAcc (pp)	ΔTok
GSM8K	512	85.42	77.08	-8.33	+5.3%
GSM8K	1024	90.63	89.58	-1.04	+12.0%
GSM8K	2048	92.71	93.75	+1.04	+15.2%
MATH-500	512	20.83	21.88	+1.04	-0.1%
MATH-500	1024	36.46	36.46	0.00	+1.6%
MATH-500	2048	54.17	57.29	+3.13	+4.5%

F Secondary Theory

The main text keeps the two results the experiments turn on. The remaining statements are collected here.

F.1 Monotonicity, and what it does not give

Lemma 5 (Bayes-risk monotonicity)**.** *If $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ then $\mathrm{I}(H;\mathcal{F}_t) \le \mathrm{I}(H;\mathcal{F}_{t+1})$, and the Bayes error $\mathcal{R}_t^\star = \mathbb{E}[\min\{R_t, 1-R_t\}]$ is non-increasing in t.*

Lemma 5 says extra prefixes cannot make the *optimal* predictor worse on average. It does not say divergence grows, that a chosen estimator improves pointwise, or that the answer becomes more correct. Since R_t is a bounded martingale, new tokens move the risk in either direction. This appendix collects the monotonicity lemma, a non-vacuous selective-stopping guarantee, a path-factorized inverse-survival correction for pruning, and a plug-in control-regret bound. Defining the ε-predictive horizon $t_\varepsilon = \inf\{t : \mathcal{R}_t^\star - \mathcal{R}_T^\star \le \varepsilon\}$ makes "does the signal arrive early?" precise but not yet measurable. The next result makes it measurable.

F.2 Safe stopping that is not vacuous

The standard confidence-sequence stopping argument controls risk but says nothing about how often the system answers, and is therefore satisfied by a rule that never answers. We state both halves.

Proposition 6 (Selective stopping, risk and efficiency)**.** *Let $r_t^{\mathrm{stop}} = \mathbb{P}(\widehat{A}_t \neq A^\star \mid \mathcal{F}_t)$ and let $(U_t)_{t\le T}$ be an adapted upper confidence sequence, $\mathbb{P}(\forall t \le T :\ r_t^{\mathrm{stop}} \le U_t) \ge 1-\delta$, with width w_t in the sense that $U_t \le r_t^{\mathrm{stop}} + w_t$ on the same event. Let $\tau = \inf\{t \le T : U_t \le \varepsilon\}$ and answer only on $\mathcal{A} = \{\tau \le T\}$. Then*

$$\mathbb{P}(\widehat{A}_\tau \neq A^\star,\ \mathcal{A}) \ \le\ \varepsilon\,\mathbb{P}(\mathcal{A}) + \delta, \qquad \mathbb{P}(\mathcal{A}) \ \ge\ \mathbb{P}\big(\exists\, t \le T :\ r_t^{\mathrm{stop}} + w_t \le \varepsilon\big) - \delta. \tag{6}$$

The second inequality makes the first meaningful. The answer rate degrades gracefully in the width of the confidence sequence, and a report of the risk bound without the realized $\mathbb{P}(\mathcal{A})$ is uninformative. Constructing a calibrated, time-uniform U_t under distribution shift is the substantive difficulty; conformal and game-theoretic tools are the natural route (Howard et al., 2021; Quach et al., 2024; Ramdas et al., 2023). Small divergence alone provides no such guarantee.

F.3 Adaptive pruning breaks the reference cohort, and how to repair it

A live controller changes its own data distribution. The initial S paths are i.i.d., but survivors of pruning are selected on their prefixes, so the unweighted survivor measure is biased for μ_t. This is invisible in offline detection and central online. We give the correction.

Assumption 1 (Randomized predictable pruning)**.** At each checkpoint the controller retains path s with probability $p_u^{(s)} \in [p_{\min}, 1]$, $p_{\min} > 0$. The probability is measurable with respect to that path's history before the retention coin is drawn; conditional on the prompt, root paths and retention coins are independent across s, and each coin is independent of future tokens given its pathwise history. This excludes bundle-coupled pruning rules.

Proposition 7 (Unbiased monitoring under pruning)**.** *Let* $K_t^{(s)} = \prod_{u \le t} \kappa_u^{(s)}$ *record survival and* $W_t^{(s)} = \prod_{u \le t} p_u^{(s)}$*. Under Assumption 1, the survival-weighted measure* $\widetilde{\mu}_t = S^{-1} \sum_s \left(K_t^{(s)}/W_t^{(s)}\right) \delta_{Z_t^{(s)}}$ *satisfies* $\mathbb{E}\langle \widetilde{\mu}_t, f\rangle = \mathbb{E}_{\mu_t}[f]$ *for every bounded* f*, and for* $f \in [0,1]$*,* $\mathbb{P}(|\langle \widetilde{\mu}_t, f\rangle - \mathbb{E}_{\mu_t} f| \ge \epsilon) \le 2\exp(-2S\epsilon^2 p_{\min}^{2t})$*. The unweighted survivor average is biased in general, with bias equal to the covariance between survival and* f *under* μ_t*, normalized by the survival probability.*

The exponent shows the cost of aggressive pruning. The effective sample size decays geometrically, so a deployed system should keep a small unpruned *reference cohort* for monitoring and prune only the *active cohort* used to produce the answer. This separation into a reference cohort, active cohort, and survival weights makes the offline detector and the online controller estimate the same quantity.

F.4 From calibration error to control regret

Proposition 8 (Plug-in control regret)**.** *Fix a finite checkpoint grid with* K *decision epochs. Let* $Q_k(a)$ *be the optimal action values of the control problem (Appendix B) including the stop action, and let* $\widehat{Q}_k$ *satisfy* $\sup_a |\widehat{Q}_k(a) - Q_k(a)| \le \varepsilon_k$ *almost surely. Then the greedy policy* $\widehat{\pi}$ *with respect to* $\widehat{Q}$ *obeys*

$$\mathbb{E}\big[J^{\widehat{\pi}}\big] - \mathbb{E}\big[J^{\star}\big] \;\le\; 2\sum_{k=0}^{K-1} \varepsilon_k. \tag{7}$$

Proposition 8 gives the link that AUROC-only evaluations miss. Token savings are governed by the *absolute* accuracy of the value estimates, not by ranking quality. Because the stop action's value is set by $\mathbb{P}(\widehat{A}_t \neq A^\star \mid \mathcal{F}_t)$, the binding ε_k is the error in predicting *correctness*, not disagreement. Section 5 reports that correctness prediction is substantially harder than same-bundle disagreement prediction; AUROC alone, however, does not instantiate the sup-norm premise of the proposition.

G Proofs

G.1 Proof of Lemma 5

Monotonicity of mutual information under nested sub-σ-fields gives $\mathrm{I}(H;\mathcal{F}_t) \le \mathrm{I}(H;\mathcal{F}_{t+1})$. For the Bayes risk, $R_t = \mathbb{E}[R_{t+1} \mid \mathcal{F}_t]$ by the tower property, and $h(p) = \min(p, 1-p)$ is concave, so conditional Jensen gives $h(R_t) \ge \mathbb{E}[h(R_{t+1}) \mid \mathcal{F}_t]$; taking expectations yields $\mathcal{R}_t^\star \ge \mathcal{R}_{t+1}^\star$. □

G.2 Proof of Theorem 1

(i) Since R is a martingale, $\mathbb{E}[R_s \mid \mathcal{F}_t] = R_t$, hence $\mathbb{E}[V_s \mid \mathcal{F}_t] = \mathbb{E}[R_s \mid \mathcal{F}_t] - \mathbb{E}[R_s^2 \mid \mathcal{F}_t] = R_t - \mathbb{E}[R_s^2 \mid \mathcal{F}_t]$, so $V_t - \mathbb{E}[V_s \mid \mathcal{F}_t] = \mathbb{E}[R_s^2 \mid \mathcal{F}_t] - R_t^2 = \mathrm{Var}(R_s \mid \mathcal{F}_t) = \mathbb{E}[(R_s - R_t)^2 \mid \mathcal{F}_t]$.

(ii) Apply (i) on each grid cell, take expectations, and telescope: $\sum_k \mathbb{E}[(R_{t_{k+1}} - R_{t_k})^2] = \mathbb{E}[V_{t_0}] - \mathbb{E}[V_{t_K}]$. At $t_K = T$ the outcome is measurable, so $R_T = H \in \{0,1\}$ and $V_T = 0$. Thus the unconditional telescope ends at $\mathbb{E}[V_0]$; conditional on a fixed prompt x, R_0 is deterministic and $V_0 = R_0(1 - R_0)$.

(iii) Write $\widehat{R}_t - H = (\widehat{R}_t - R_t) + (R_t - H)$. The cross term vanishes because $\widehat{R}_t - R_t$ is $\mathcal{F}_t$-measurable and $\mathbb{E}[R_t - H \mid \mathcal{F}_t] = 0$, giving $B_t = \mathbb{E}[(R_t - H)^2] + \mathbb{E}[(\widehat{R}_t - R_t)^2]$. Moreover $\mathbb{E}[(R_t - H)^2] = \mathbb{E}\big[\mathbb{E}[(R_t - H)^2 \mid \mathcal{F}_t]\big] = \mathbb{E}[V_t]$ because H is conditionally Bernoulli with mean R_t. Differencing at t and s and applying (i) gives Eq. (2). □

G.3 Proof of Theorem 2

Let $Z_i = \mathbf{1}\{Y_i = b\} - \mathbf{1}\{Y_i = a\}$, so $Z_i \in \{-1, 0, 1\}$ with probabilities u, $1 - u - v$, and v. For $u > v$, the event $N_b \ge N_a$ is $S^{-1}\sum_i Z_i \ge 0$, a right-tail deviation from mean $v - u < 0$. Its log moment-generating function is $\Lambda(\lambda) = \log(ve^{\lambda} + ue^{-\lambda} + 1 - u - v)$, minimized at $e^{2\lambda^\star} = u/v$.

Cramér's theorem gives exponent $-\Lambda(\lambda^\star) = -\log(1-(\sqrt{u}-\sqrt{v})^2)$. The reverse statement follows by exchanging u and v. □

G.4 Proof of Corollary 3

For a finite alphabet with a unique correct mode, plurality error is contained in the union over incorrect classes of pairwise inversion events, each exponentially rare by Theorem 2. If an incorrect class has larger mass, correct plurality requires the reverse rare event against that class. At equality the exponent vanishes and the limiting probability depends on the joint central-limit behavior and tie-breaking. □

G.5 Proof of Proposition 6

Let $\mathcal{G} = \{\forall t \le T : r_t^{\mathrm{stop}} \le U_t\}$, so $\mathbb{P}(\mathcal{G}) \ge 1-\delta$. On $\mathcal{A}\cap\mathcal{G}$, evaluating at the stopping time gives $r_\tau^{\mathrm{stop}} \le U_\tau \le \varepsilon$; on $\mathcal{G}^c$ bound the joint probability by $\mathbb{P}(\mathcal{G}^c)$. Hence $\mathbb{P}(\widehat{A}_\tau \ne A^\star, \mathcal{A}) = \mathbb{E}[\mathbf{1}_{\mathcal{A}} r_\tau^{\mathrm{stop}}] \le \varepsilon\mathbb{P}(\mathcal{A}\cap\mathcal{G}) + \delta \le \varepsilon\mathbb{P}(\mathcal{A}) + \delta$. For the efficiency half, on $\mathcal{G}$ the width condition gives $U_t \le r_t^{\mathrm{stop}} + w_t$, so $\{\exists t \le T : r_t^{\mathrm{stop}} + w_t \le \varepsilon\} \cap \mathcal{G} \subseteq \{\exists t : U_t \le \varepsilon\} = \mathcal{A}$. Subtracting $\mathbb{P}(\mathcal{G}^c) \le \delta$ gives the stated lower bound on $\mathbb{P}(\mathcal{A})$. □

G.6 Proof of Proposition 7

Let $\mathcal{H}_t = \sigma(\Gamma_{\le t}^{(1:S)})$ be generated by the uncensored trajectories. Fix s and expand $M_t^{(s)} = \prod_{u\le t} \kappa_u^{(s)}/p_u^{(s)}$. Condition on $\mathcal{H}_t$ and process the factors in increasing u. By Assumption 1, $p_u^{(s)}$ is measurable with respect to $\mathcal{H}_t \vee \sigma(\text{coins before } u)$, and the coin at u is conditionally Bernoulli$(p_u^{(s)})$ given that field and independent of later tokens. Hence $\mathbb{E}[\kappa_u^{(s)}/p_u^{(s)} \mid \mathcal{H}_t \vee \sigma(\text{coins before } u)] = 1$, and by the tower property $\mathbb{E}[M_t^{(s)} \mid \mathcal{H}_t] = 1$. Since $Z_t^{(s)}$ is $\mathcal{H}_t$-measurable, $\mathbb{E}[M_t^{(s)} f(Z_t^{(s)})] = \mathbb{E}[f(Z_t^{(s)})\mathbb{E}[M_t^{(s)} \mid \mathcal{H}_t]] = \mathbb{E}_{\mu_t}[f]$; averaging over s gives unbiasedness. Each summand lies in $[0, p_{\min}^{-t}]$ for $f \in [0,1]$, and the S paths with their coins are independent, so Hoeffding's inequality gives the deviation bound after rescaling. For the unweighted survivor average, $\mathbb{E}[K_t^{(s)} f(Z_t^{(s)})]/\mathbb{E}[K_t^{(s)}] = \mathbb{E}_{\mu_t}[f] + \mathrm{Cov}(K_t^{(s)}, f(Z_t^{(s)}))/\mathbb{E}[K_t^{(s)}]$, which is nonzero whenever survival is informative about f. □

Remark 9. The exponent $p_{\min}^{2t}$ is pessimistic. It reflects the worst case in which every path is at risk of pruning at every checkpoint. In practice pruning happens at few checkpoints, and the relevant quantity is the effective sample size $(\sum_s M_t^{(s)})^2/\sum_s (M_t^{(s)})^2$, which we recommend logging online. A self-normalized empirical-Bernstein version with data-dependent weights is left open.

G.7 Proof of Proposition 8

Backward induction on the checkpoint index. Let $J_k^\star$ and $J_k^{\widehat{\pi}}$ be the optimal and greedy value functions from epoch k onward and $\Delta_k = \sup|J_k^{\widehat{\pi}} - J_k^\star|$. At the terminal epoch $\Delta_K = 0$. At epoch k, let $a^\star$ be optimal and $\widehat{a}$ greedy for $\widehat{Q}_k$. Then $Q_k(\widehat{a}) \le \widehat{Q}_k(\widehat{a}) + \varepsilon_k \le \widehat{Q}_k(a^\star) + \varepsilon_k \le Q_k(a^\star) + 2\varepsilon_k$, so the one-step action choice costs at most $2\varepsilon_k$ while continuation adds at most Δ_{k+1}. Hence $\Delta_k \le 2\varepsilon_k + \Delta_{k+1}$, and unrolling gives $\Delta_0 \le 2\sum_{k<K}\varepsilon_k$. Taking expectations yields Eq. (7). □

H Semantic Dispersion

The active-only analysis found larger semantic second moments for trajectories that ended in error. We report the descriptive result with an important qualification. Completion creates different survivor cohorts, and the ratio is strongly anti-correlated with the number of surviving correct prompts. A subsequent length-matched pilot further weakens the interpretation. On TriviaQA, active-only error/correct ratios are approximately 1.03-1.35 at 32-64 tokens and collapse to 1.00 by 128 tokens for the tested Llama-1B and Gemma-4B models; on Qwen-0.6B/GSM8K the intervals are extremely wide and change direction. Ratios computed over all paths can be enormous because completed paths act as distinct absorbing states, confirming that EOS/completion geometry dominates that estimand.

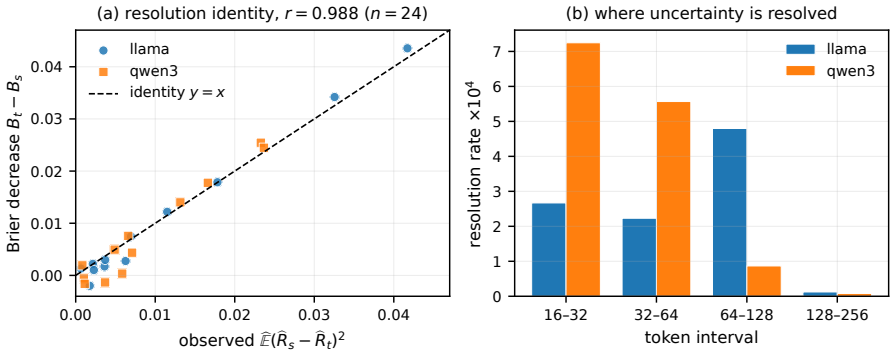


Figure 10: Predictor diagnostic motivated by Theorem 1 ($S = 16$, two 8B models, three datasets, four intervals; $n = 24$ cells). The horizontal quantity uses squared increments of the learned predictor, not the latent posterior, so the identity line is a visual reference rather than a theorem test. Panel (b) reports predictor movement per token.

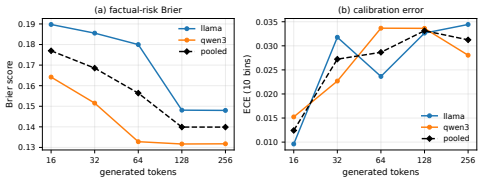


Figure 11: Factual-risk Brier score and ECE pooled over two 8B models and three datasets. Brier improves to 128 tokens and then changes little; ECE remains below 0.034.

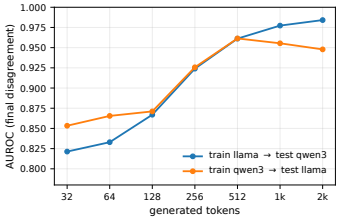


Figure 12: Cross-family ranking transfer for final same-bundle disagreement. This does not measure calibration or independent-pool disagreement.

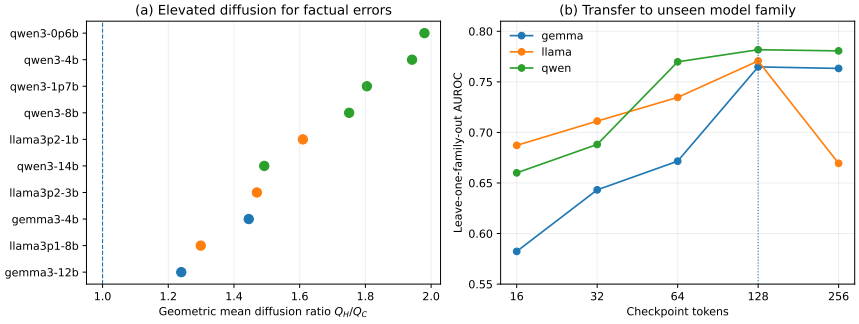


Figure 13: Active-only dispersion summary and leave-one-family-out ranking result. The dispersion panel is confounded by differential completion and is not used in the main claims.

We therefore present semantic dispersion only as an unresolved exploratory signal, not evidence for a diffusion or a robust correctness feature.

I Cross-Family Transfer

What we look at. Whether a predictor fitted on two model families preserves ranking performance on a third it has never seen. This evaluates ranking transfer only; it does not establish calibration transfer or a model-agnostic controller.

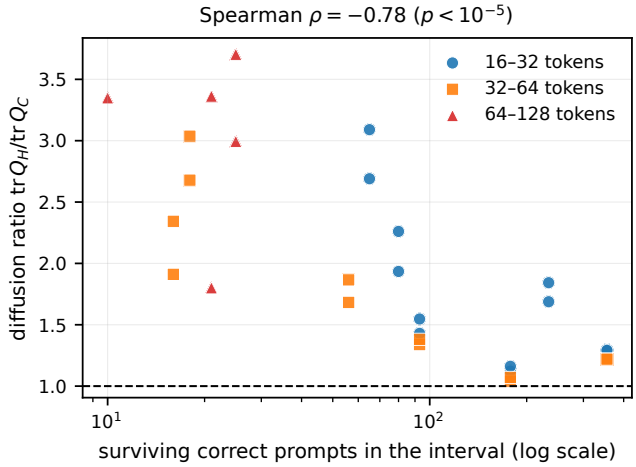


Figure 14: Attrition diagnostic. The largest dispersion ratios occur where the surviving correct group is most depleted. Length-matched pilots do not establish a stable active-only effect.

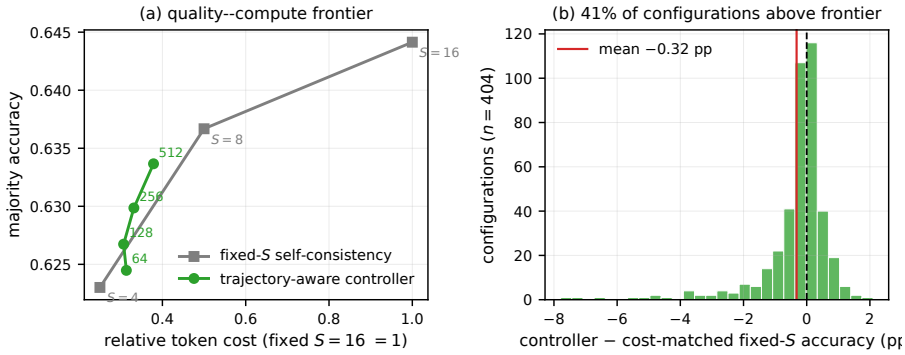


Figure 15: Offline policy replay against a pre-generated pool. (a) Cross-family weighted frontier over 39,428 model-question evaluations; the controller lies near the fixed-S interpolation. (b) Difference from cost-matched fixed-S over 404 configurations. Mean -0.32 pp, median -0.06 pp, 41% positive.

What we find. A leave-one-family-out predictor trained on two families and evaluated on the third reaches AUROC 0.765 on unseen Gemma, 0.771 on unseen Llama and 0.782 on unseen Qwen at 128 tokens (Appendix Figure 13b). Llama is the exception and decreases to 0.669 by 256. Completion-driven selection is a plausible explanation, not an established cause. A family-dependent fraction of paths has terminated by then, so late feature distributions are not directly comparable. The path-factorized result in Proposition 7 does not cover arbitrary bundle-coupled selection. On the reasoning suite the corresponding transfer is stronger and monotone (Appendix Figure 12). It reaches 0.869 at 128 and 0.961 at 512.

J Exploratory Decoder Sweep

What we look at. Whether raising token-level temperature on diagnostic paths improves the measured predictor signal.

What we find. With $T_{\text{answer}} = 0.7$ fixed, AUROC at 128 tokens increases with probe temperature in the reported decoder sweep. For 16 probe paths, it rises from 0.726 at $T_{\text{probe}} = 0.3$ to 0.813 at 1.0. Eight paths at $T = 1.0$ also exceed sixteen at $T = 0.7$ on that observability metric. This is an empirical decoder intervention, not a test of an answer-level tempering theorem. Autoregressive token temperature, top-p/top-k, sequence length, and EOS probabilities can change the induced

answer-class law and its support, so we make no invariant-ordering or limiting-error claim. Probe paths also cannot be reused in the answer ensemble, and the mixed-temperature controller did not dominate after token accounting.

K Limitations and Scope

External and token-axis baselines. The s1-style budget-forcing control in Appendix E covers a reasoning-length intervention and shows no consistent Pareto improvement. The PRM experiment spans Qwen3-8B and Llama-3.1-8B generators. Net of a free length-heuristic control on the identical candidate set, no cell of twenty has an interval excluding zero, so we do not claim an established fixed-candidate selection gain. Its efficiency claim is narrower still. Padded micro-batches do not reproduce the reference rewards in the tested implementation, and a separate online Qwen latency benchmark does not confirm the favorable offline $S = 2$ accounting. A second independent verifier and a validated batched serving implementation would be needed for a general verifier-efficiency claim. The DeepConf-style replay is a useful selection-side diagnostic, but it is not a faithful reproduction of the official method.

Statistics and scope. Learned-prediction intervals use paired question bootstrap and cross-validation groups by underlying question. High-S curves use question-bootstrap intervals. Stopping uses 100 reveal orders for ordering robustness and a separate nested 2,000-replicate question bootstrap for population uncertainty; the fixed-S frontier is recomputed inside every replicate. The broad pooled historical curves still summarize repeated model-question evaluations and should not be read as independent questions. Regime labels based on intervals for Δ_x depend on estimation-pool size, so the main text relies on C_S, A_S, R_S rather than category proportions.

Compute accounting. Stopping tables report both generation-only and generation-plus-policy-CPU cost. The measured posterior/ESC decision overhead is sub-millisecond and does not materially change the matched frontier on the evaluated offline replay. This still does not model a production serving stack with queueing, dynamic batching, or memory contention. PRM accounting remains more implementation-sensitive because batch-> 1 scoring did not reproduce the reference rewards; we therefore make no deployment-level verifier-efficiency claim.

Representation, answer measurement, and length. Prefix NLL often matches or exceeds richer semantic trajectory features, so semantic dispersion is secondary rather than a headline mechanism. Factual canonicalization is also part of the measurement process. The harmful-sampling pattern that the exposed matcher produces on Llama/TriviaQA does not appear under the robust matcher, and no canonicalizer built on the robust normalizer reproduces it. Generation length is a separate confound. Longer Qwen/GSM8K and ARC controls reduce cap hits to 4.6% and 1.1% and materially shrink the earlier Qwen/GSM stopping gain. The original Qwen/TriviaQA pool has 63.8% cap hits and is excluded from headline intervention claims. Llama/MATH-500 remains length-sensitive at 4096 tokens (12.6% cap hits), so its intervention results are secondary.

Reproducibility. Appendix M records model revisions, decoding parameters, seed policy, prompt and trajectory feature classes, answer extraction/canonicalization, A/B splits, fair-tie plurality, high-S, stopping bootstrap, and PRM scoring. The generator revisions recovered from the run cache are `b968826d9c46dd6066d109eabc6255188de91218` for Qwen3-8B and `0e9e39f249a16976918f6564b8830bc894c89659` for Llama-3.1-8B-Instruct. The PRM revision is pinned separately below.

Information versus computation. Conditional on the full (X, θ, ψ), independent pools are independent, so the independent-pool experiment cannot establish new Shannon information beyond the prompt. Its claim is explicitly observer-relative. Generated prefixes are Monte-Carlo computations that improve a particular finite predictor. A stronger prompt representation or hypothesis class may reduce this increment. Likewise, high AUROC for disagreement does not imply calibrated correctness or positive control value; the negative learned-controller result is evidence for this gap rather than an algorithmic success claim.

L External Replication on EVOUNA

EVOUNA (Wang et al., 2023a) contains human correctness judgments for answers from five Open-QA systems, namely DPR+FiD, GPT-3.5, ChatGPT-3.5, GPT-4 and BingChat, on Natural Questions (3,020 questions after dropping those the authors flag as improper) and TriviaQA (1,938). It gives this paper a reference produced by other people, on systems we did not run, annotated before this work existed. We use it to check that our matching rules behave as claimed on external answers, and to test the interaction law with correctness taken from human labels rather than from a permissive proxy rule.

Table 4: Accuracy on EVOUNA under human judgment and under three matching rules. The long-form LLM systems are unmeasurable by exact match. GPT-4 scores 0.000 on NQ against a human accuracy of 0.788. Only DPR+FiD, which emits short spans, is within reach of a strict rule. This is a property of the systems EVOUNA covers, not of the matchers alone. Our SQuAD-official column differs slightly from the exact-match figures reported by Wang et al. (2023a), for example 0.561 against their 0.592 on NQ/DPR+FiD, because their exact-match implementation is not the SQuAD-official normalizer.

dataset	system	human	SQuAD official	plain exact	robust matcher
NaturalQuestions	DPR+FiD	0.690	0.561	0.546	0.605
NaturalQuestions	GPT-3.5	0.655	0.028	0.021	0.034
NaturalQuestions	ChatGPT-3.5	0.730	0.015	0.004	0.019
NaturalQuestions	GPT-4	0.788	0.000	0.000	0.004
NaturalQuestions	BingChat	0.799	0.000	0.000	0.000
TriviaQA	DPR+FiD	0.815	0.717	0.705	0.745
TriviaQA	GPT-3.5	0.784	0.229	0.099	0.258
TriviaQA	ChatGPT-3.5	0.844	0.075	0.024	0.083
TriviaQA	GPT-4	0.902	0.038	0.001	0.080
TriviaQA	BingChat	0.896	0.000	0.000	0.000

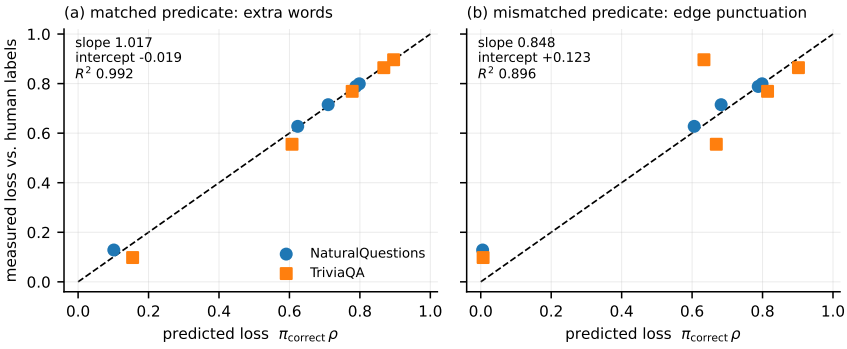


Figure 16: The interaction law on EVOUNA, with correctness taken from human labels. In panel (a) the predicate matches the defect that SQuAD normalization is exposed to, namely the answer being longer than the shortest gold alias, and the ten cells lie on the identity line. In panel (b) the edge-punctuation predicate describes a defect that SQuAD normalization removes, the fit degrades, and it acquires a positive intercept. Predictor support is narrow in both panels, since most cells sit above 0.6, where $\pi_{\text{correct}}\rho$ approaches ρ.

What the table shows. Exact-match evaluation is simply inapplicable to the long-form systems EVOUNA covers. Four of the ten cells score below 0.001 under every strict rule while humans score them between 0.73 and 0.90. Realistic pipelines extract a short span before matching, as ours does, and the raw comparison here should be read as an upper bound on the damage rather than as a description of how these systems are normally scored. The informative variation for our purposes comes from the contrast between DPR+FiD, whose short spans keep it within reach of a strict rule, and the rest.

What the figure shows, and what it does not. With the matched predicate the fit is slope 1.017, intercept -0.019, $R^2 = 0.992$, with a leave-one-out slope range of $[0.971, 1.083]$. Accuracy alone gives $R^2 = 0.21$. With the mismatched predicate the slope falls to 0.848 and the intercept rises to $+0.123$, significantly away from zero. We report the second result because it is the specificity control. A law that fitted equally well under any plausible predicate would be evidence of arithmetic rather than of mechanism.

The design limitation is real and we do not minimise it. The style rate among correct answers takes values 0.147 and 0.190 for DPR+FiD and between 0.77 and 1.00 for everything else, so the largest gap between consecutive cells is 0.584 and the regression has roughly two effective clusters rather than ten independent points. The released script prints this diagnostic and the accompanying warning by default.

M Reproducibility Protocol

This section gives implementation-level pseudocode for the experiments underlying the main claims. It also makes explicit which quantities are posterior statements about one sampled bundle and which are observer-relative Monte-Carlo probe experiments.

M.1 Shared generation and answer measurement

The headline generators are `Qwen/Qwen3-8B`, revision `b968826d9c46dd6066d109eabc6255188de91218`, and `meta-llama/Llama-3.1-8B-Instruct`, revision `0e9e39f249a16976918f6564b8830bc894c89659`. Decoding uses random seed 20260824, temperature 0.7, top-$p = 1.0$, top-$k = -1$, min-$p = 0$, and maximum model length 8192. Qwen3 uses thinking mode. The original reasoning pools use at most 2048 generated tokens and factual pools 256. Because the truncation audit found material cap hits, we additionally generate 4096-token controls for Qwen/GSM8K, Qwen/ARC, and Llama/MATH-500. The main intervention claims use the clean or length-controlled slices described in Section 5. Each independent-prediction pool contains 64 root samples per question split into fixed 32-sample A/B halves.

Responses are parsed by taking the last explicit `FINAL:` field when present, then the last `boxed{}` expression, and finally the last nonempty line. Factual strings are Unicode-NFKC normalized, lower-cased, stripped of English articles and most punctuation, and whitespace-normalized. A factual response matching a dataset alias exactly, or by the stated short suffix rule, maps to `__CORRECT__`; other factual answers use their normalized surface form. Math correctness uses `math_verify` when available with numeric fallback. MCQ answers are canonical option letters. Empty extractions map to `__UNPARSED__`. Fair plurality evaluation assigns a tied correct class probability $1/k$ when k classes share the maximum count; it never uses lexical class order.

M.2 Finite-observer independent-pool prediction

The robustness analysis compares several sample-free prompt observers. Text baselines use `sentence-transformers/all-MiniLM-L6-v2`; stronger baselines add generator-side prompt statistics or a frozen prompt representation from the generator itself. All learned variants use grouped cross-validation by question. Trajectory models augment the prompt representation with A-pool prefix statistics only; B is reserved for the independent disagreement/error target. Cheap trajectory features include checkpoint, mean/std sampled-token prefix NLL, completion fraction, provisional-answer fraction, and provisional-answer entropy. Full features add semantic dispersion and semantic-cluster summaries. Paired question bootstrap is used for AUROC differences, and the A/B sensitivity analysis repeats the experiment over 20 random 32/32 repartitions. Threshold sensitivity reports $\gamma \in \{0.20, 0.30, 0.35, 0.40, 0.50\}$.

The same scalar evaluation is repeated on completion-fraction strata, including the subset with no completed A-paths. Because $A \perp B \mid X, \theta, \psi$, Algorithm 1 measures observer-relative predictive value; it is not an estimator of $I(H_B; \mathcal{F}_t^A \mid X, \theta, \psi)$, which is zero under the fixed-law model.

Algorithm 1 Independent-pool computational-probe evaluation

Require: Questions $\mathcal{D}$, disjoint root-sample pools A, B, checkpoints $\mathcal{T}$

for all $t \in \mathcal{T}$ **do**
  for all $x \in \mathcal{D}$ **do**
    $z_x^{\text{prompt}} \leftarrow \text{PromptFeatures}(x)$
    $z_{x,t}^{\text{cheap}} \leftarrow \text{CheapFeatures}(A_{\le t}(x))$
    $z_{x,t}^{\text{full}} \leftarrow \text{FullFeatures}(A_{\le t}(x))$
    $z_{x,t}^{\text{nll}} \leftarrow \text{MeanPrefixNLL}(A_{\le t}(x))$
    $\widehat{\nu}_B(x) \leftarrow \text{AnswerHistogram}(B(x))$
    $y_x^{\text{dis}} \leftarrow \mathbf{1}\{U(\widehat{\nu}_B(x)) > 0.35\}$
    $y_x^{\text{err}} \leftarrow \mathbf{1}\{\text{Plurality}(B(x)) \neq A_x^\star\}$
  end for
  for all $z \in \{z^{\text{prompt}}, z^{\text{cheap}}, z^{\text{full}}\}$ **do**
    fit/evaluate the same grouped logistic protocol and record held-out AUROC
  end for
  evaluate prefix NLL as a scalar baseline against the same B-only targets
end for

Algorithm 2 Finite-budget acquisition-selection decomposition with fair ties

Require: Completed B pool with 32 samples per question

for all $S \in \{1, 2, 4, 8, 16, 32\}$ **do**
  for all question x **do**
    $Y_S(x) \leftarrow$ first S canonical answer classes in B
    $c_x \leftarrow \mathbf{1}\{A_x^\star \notin Y_S(x)\}$
    let k_x be the number of classes tied for maximum count in $Y_S(x)$
    $p_x^{\text{corr}} \leftarrow 1/k_x$ if $A_x^\star$ is among the tied maxima, else 0
    $r_x \leftarrow 1 - p_x^{\text{corr}}$; $a_x \leftarrow 0$ if $c_x = 1$, else r_x
  end for
  $C_S \leftarrow \text{mean}_x(c_x)$; $A_S \leftarrow \text{mean}_x(a_x)$; $R_S \leftarrow \text{mean}_x(r_x)$
  bootstrap question IDs for uncertainty bands
end for

M.3 Finite-budget acquisition-selection curves

For $S \in \{1, 2, 4, 8, 16, 32\}$, evaluation uses the first S candidates of the fixed B pool. For each question, $N_S^\star$ counts correct candidates and $\widehat{A}_S$ is plurality.

No $N_S^\star = 0$ observation is interpreted as structural zero support.

M.4 Random-order posterior-consistency stopping

For each question, the 32 B candidates are permuted independently for each of 100 reveal-order repetitions. Fixed-S and adaptive policies use the same permutation in a repetition. ESC uses unanimity of the latest window. The primary posterior-consistency heuristic is the rule stated in Section 5. It uses symmetric Dirichlet parameter $\alpha = 0.5$, observed classes plus one zero-count "other" class, 3,000 posterior draws, a minimum of 4 samples, a maximum of 32, and threshold 0.90. The policy does not stop while the observed plurality leader is tied. Accuracy after stopping uses uniform random tie handling, so if the correct class is one of k classes tied for maximum count it contributes $1/k$ expected correctness.

For fixed policies with measured generation costs c_i and accuracies a_i, the reference at adaptive cost c is the upper randomized envelope

$$F(c) = \max_{i,j,\lambda \in [0,1]:\ \lambda c_i + (1-\lambda) c_j \le c} \lambda a_i + (1-\lambda) a_j.$$

Algorithm 3 Stopping replay with question-level uncertainty

Require: 32-candidate B pools, policy π, 100 reveal-order seeds, 2,000 bootstrap seeds

for $r = 1, \ldots, 100$ **do**
  for all question x **do**
    randomly permute its 32 candidates
    reveal candidates until π stops or $S = 32$
    record fair-tie expected correctness and generation/policy cost
  end for
  compute adaptive and fixed-S points on the same permutations
end for
for $b = 1, \ldots, 2000$ **do**
  resample question IDs with replacement
  recompute the adaptive point and randomized fixed-S frontier
  average the frontier difference over reveal orders within replicate b
end for
report question-bootstrap 95% intervals; report reveal-order ranges separately

Algorithm 4 PRM selection on an identical candidate set

Require: Candidate traces $\{\Gamma_i(x)\}_{i=1}^{S}$, frozen score summary g

for all question x **do**
  for $i = 1, \ldots, S$ **do**
    obtain step rewards $r_{i1}, \ldots, r_{im_i}$ from the PRM
    $w_i \leftarrow g(r_{i1}, \ldots, r_{im_i})$
  end for
  aggregate weights by canonical answer class and emit the largest-weight class
  record majority/PRM error, non-coverage, generation time, and PRM time
end for
bootstrap questions for the paired same-candidate accuracy difference
compare offline total time with the randomized fixed-S envelope

M.5 Process-reward selection on a fixed candidate set

The verifier is `Qwen/Qwen2.5-Math-PRM-7B`, revision `0610740060112df12585d00a1c5f4624d2f59051`. Candidate generation is unchanged. Reasoning traces are segmented into steps; the PRM reward probability is read at the model's step separator positions. Candidate summaries considered on Qwen calibration include minimum, mean, last-step, and product/log-product variants; the selected statistic is frozen before held-out and Llama transfer evaluation. Weighted voting emits

$$\widehat{A}_{\mathrm{PRM}} = \arg\max_{a} \sum_{i=1}^{S} w_i \mathbf{1}\{A_i = a\}.$$

Because the candidate set is identical to majority voting, C_S is identical by construction and only A_S can change.

Padded microbatches 2, 4, and 8 are used for serving only if they reproduce unpadded batch-1 rewards within the validation tolerance. They failed this equivalence test in the reported runs, so final PRM scores and timings use batch 1.

M.6 Stop-then-Verify and condition mixing

Stop-then-Verify first applies a whole-sample stopping policy and then invokes the frozen PRM only when a calibration-selected vote-margin ambiguity threshold is exceeded. Its final point is compared with an envelope containing fixed majority, fixed PRM, stopping-only, and always-PRM baselines. The negative result is interpreted only for this threshold router.

For prompt-condition mixing, the total root-sample budget is held fixed. If condition k receives n_k samples with $\sum_k n_k = S$, all S canonical answers are pooled and passed to the same plurality rule as the single-condition control; paired differences are computed by question.

M.7 Released robustness artifacts

The saved artifacts record question IDs, A/B assignments, generation token counts/times, feature checkpoints, 20 random A/B repartitions, disagreement-threshold sensitivity, alternative TriviaQA canonicalizers, fair-tie high-S curves, 100 stopping reveal orders, 2,000 nested question-bootstrap replicates, fixed-S frontiers, PRM revision and scoring statistics, and the 4096-token length controls. Immutable generator revisions are recorded in the shared-generation subsection. Together, these artifacts permit recomputation of the headline robustness tables from saved generations without rerunning the language models.