TRAJUQ: PREDICTING LLM REASONING FAILURE UNDER LIMITED OBSERVATION BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Short continuations offer evidence of failure risk before a large language model (LLM) answers, but observing more branches or following them longer consumes tokens. Can better readout of finite, imperfect probe scores compensate for fewer observations? We address this question with TrajUQ, combining the endpoint mean and the empirical distribution of centered branch-score vectors to predict continuation failure. We evaluate 540 questions across three tasks and thirteen model configurations using question-grouped out-of-fold prediction. With ten-branch, 64-token fits frozen, our TrajUQ readout improves Brier over the learned mean by 3.125% on average across 20 matched combinations of branch count and observation depth, each using at least two branches. In our post-hoc scan, two-branch TrajUQ achieves lower Brier than the ten-branch, 64-token mean readout using 80% fewer observation tokens. At the reference budget, our matched tuning yields Brier 0.1947 versus 0.2011 for the learned mean, also improving on the tested second-order and distribution-only controls. Using these predictions, our offline screening selects prefixes with lower empirical failure at matched coverage. Our offline TrajUQ candidate selection reduces final-answer failure from 56.47% for a single answer chosen uniformly from the same candidate pool to 51.92% without updating LLM parameters. These results connect better use of limited observations to improved risk prediction and more reliable offline answer selection.

1 INTRODUCTION

Before a large language model (LLM) produces a final answer, the same reasoning prefix can lead to different outcomes. For a fixed model and decoding policy, prefix failure risk is the probability that a continuation from that prefix fails to produce a correct answer (Park et al., 2025). Short continuations sampled from that prefix provide partial evidence about this risk. Observing more branches or following them longer consumes additional tokens. Can a better readout yield more accurate risk predictions from fewer observed branches?

We obtain predictive evidence by resampling short branches from a common prefix. Prior work uses such resampling to examine possible futures (Macar et al., 2026) and short branches to guide subsequent generation (Zhu et al., 2026). Our probe, trained on individual branch outcomes, scores states from branches truncated before the final answer. We combine these scores with a separate readout to estimate prefix failure risk. Branch count and observation depth control the available evidence; the readout determines how that evidence becomes a risk estimate.

Averaging provides a principled starting point for this risk readout. Math-Shepherd's soft estimator averages success indicators from sampled completions (Wang et al., 2024a); under the policy-value formulation of future success (Setlur et al., 2025), ideal conditional failure scores also recover prefix risk in expectation. In practice, we work with finite collections of imperfect probe scores. A learned readout of the endpoint mean can calibrate the overall score level, but collections with the same mean remain indistinguishable to it even when their branch variation differs. We test whether retaining this variation improves prediction beyond a learned mean, giving both readouts comparable fitting and selection opportunities.

To test this possibility, we represent score level and branch variation jointly in TrajUQ. The endpoint mean O captures the level; centering scores across branches at each depth yields vectors whose empirical distribution is U. Building on distribution regression (Szabó et al., 2016), our joint readout

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.

learns how O and U relate to prefix risk, supervised by independently sampled full-continuation outcomes. This allows the predictive role of branch variation to depend on score level.

For models with known architectures and accessible hidden representations, we fit the probe and risk readout using labeled continuation outcomes from the target task, while keeping the LLM parameters fixed. Our contributions are:

- **Joint risk readout.** In TrajUQ, we combine endpoint mean O and centered empirical distribution U to predict prefix failure risk from score level and branch variation.
- **Better prediction at the same or lower observation cost.** With the ten-branch, 64-token predictors held fixed, we reduce Brier with TrajUQ by 3.125% over the learned mean, averaged across 20 combinations of branch count and observation depth using at least two branches. In this scan, two-branch TrajUQ outperforms the ten-branch mean with 80% fewer observation tokens (Figure 2; Table 22). At the reference budget, matched tuning also yields lower Brier for TrajUQ than for the tested alternatives (Table 1).
- **From risk prediction to answer selection.** Our offline screening selects lower-risk prefixes at matched coverage (Figure 3). Across 180 TriviaQA questions and thirteen model configurations, our selection among ten saved candidates reduces failure from 56.47% for a single answer to 51.92% (Figure 4).

From risk prediction to answer selection, TrajUQ uses limited observations to support more reliable offline selection among saved candidate answers, while keeping LLM parameters fixed.

2 Related Work

Prefix risk and process value. Prefix-risk estimation extends LLM correctness prediction (Kadavath et al., 2022) to unfinished reasoning. Math-Shepherd uses completion success proportions as soft labels (Wang et al., 2024a), while Park et al. calibrate process reward models using independently sampled continuations (Park et al., 2025). Rewarding Progress formalizes expected terminal correctness under a specified rollout policy (Setlur et al., 2025). Related targets include PUM's solve-rate gains over solving from scratch across student policies (Zhou et al., 2026) and ProbGuard's completed-response safety risk, estimated from prefix token output distributions with Monte Carlo continuation supervision (Huang et al., 2026). BetaPRM models Beta beliefs over prefix success from continuation counts (Li et al., 2026); Distributional Process Reward Models estimate conditional quantiles of PRM-estimated success probabilities (Ma et al., 2026). Our TrajUQ readout predicts absolute prefix failure risk under a fixed policy from the empirical distribution of observed branch scores.

Observing alternative reasoning continuations. Reasoning uncertainty can be examined along a single trajectory or through alternative continuations. Tracing Uncertainty uses token-level uncertainty profiles to predict the final answer's correctness (Grünefeld et al., 2026). Thought Branches resamples from a shared prefix to analyze possible futures, reasoning steps, and intervention effects (Macar et al., 2026). GUARD uses mean token entropy to select short branches at high-uncertainty transitions and redirect generation (Zhu et al., 2026). With limited branch count and observation depth, our TrajUQ readout uses short-continuation scores from a common prefix to predict its failure risk under a fixed policy.

Distribution representations and learned risk readouts. Sampled responses support both semantic and internal-state representations of uncertainty. Semantic entropy groups semantically equivalent answers, while kernel language entropy retains graded semantic similarities (Kuhn et al., 2023; Nikitin et al., 2024). INSIDE uses EigenScore to summarize the covariance spectrum of multiple response embeddings (Chen et al., 2024). Learned readouts include supervised score combinations (Chauhan et al., 2026), probes predicting sampled semantic entropy from a single generation's hidden states (Kossen et al., 2024), and SIVR's sequential use of token-level, cross-layer dispersion features for hallucination detection (Srey et al., 2026). Distribution regression provides a framework for learning from distributions represented by finite sample sets (Szabó et al., 2016). We apply distribution regression to short-continuation scores in TrajUQ, combining the endpoint mean with the empirical distribution of same-branch score vectors centered across branches at each selected depth.

Observation budgets and computation. Snell et al. show that the effectiveness of test-time computation depends on allocation and prompt difficulty (Snell et al., 2025). Dynamic feature selection studies the acquisition of informative observations under a limited budget (Gadgil et al., 2024). With reference fits frozen, we examine how short-continuation risk readouts affect predictive loss across observation-token budgets, including within-reader branch–depth allocations under a common token ceiling.

3 Problem Setup

3.1 Risk target and independent supervision

Let x denote a question together with its current reasoning prefix, and let π denote the fixed language model and decoding policy. A fresh continuation sampled from π given x has outcome $Y \in \{0, 1\}$, where $Y = 1$ indicates an incorrect final answer according to the task evaluator. Following the prefix-conditioned success probability studied by Park et al. (2025), we define its complementary *prefix failure risk* as

$$\rho(x) = \Pr_{\pi}(Y = 1 \mid x). \tag{1}$$

To supervise $\rho(x)$, we sample n independent full continuations from x under π. These label continuations are sampled independently of the short-continuation inputs, conditional on x and π. Their failure indicators $Y_1, \ldots, Y_n$ give the empirical failure rate

$$r = \frac{1}{n} \sum_{j=1}^{n} Y_j.$$

This Monte Carlo estimate provides noisy supervision for the prefix risk readout. We evaluate predicted risks against the individual continuation outcomes.

3.2 Short-continuation observations and evaluation

For prediction, we observe up to H generated tokens on each of K continuations sampled from x under π. Our task prompts require the final answer on the last line, introduced by `Final Answer:` (Appendix B.1). We observe the generated reasoning segment before this marker, excluding the marker and final-answer text. We cap each requested depth at the number of tokens in that segment (Appendix B.2).

During training, we supervise the probe with observation-branch outcomes and the prefix risk readout with independent full-continuation failure rates r. At prediction time, the fitted probe scores the bounded pre-answer branch states, and the readout uses only these scores to estimate $\rho(x)$.

Our evaluation records identify the model configuration, original question, and prefix position. We keep all records and continuations from the same original question together in data splits. Confidence intervals for predictive performance use bootstrap resampling of these question groups (Appendices B.6 and C.2).

4 Framework

4.1 From short continuations to level and structure

In TrajUQ, we represent short continuations by endpoint score level and their score-vector distribution (Figure 1). For each of the K branches sampled from context x, let $Z_{b,t}$ be its available hidden state at requested depth t. Our probe g, fitted on training-branch outcomes and frozen, produces a score $p_{b,t} = g(Z_{b,t}) \in [0, 1]$ as a proxy for eventual branch failure. We apply the same probe at the observation horizon H and at selected depths S within that horizon. We use scores at H for the endpoint mean and at S for the distribution; S need not include H. Probe implementation and endpoint selection for truncated branches are detailed in Appendix B.

We represent score level by the endpoint mean $O = \bar{p}_H = K^{-1} \sum_{b=1}^{K} p_{b,H}$. Averaging has precedent in Math-Shepherd's soft estimator of completion success (Wang et al., 2024a). For

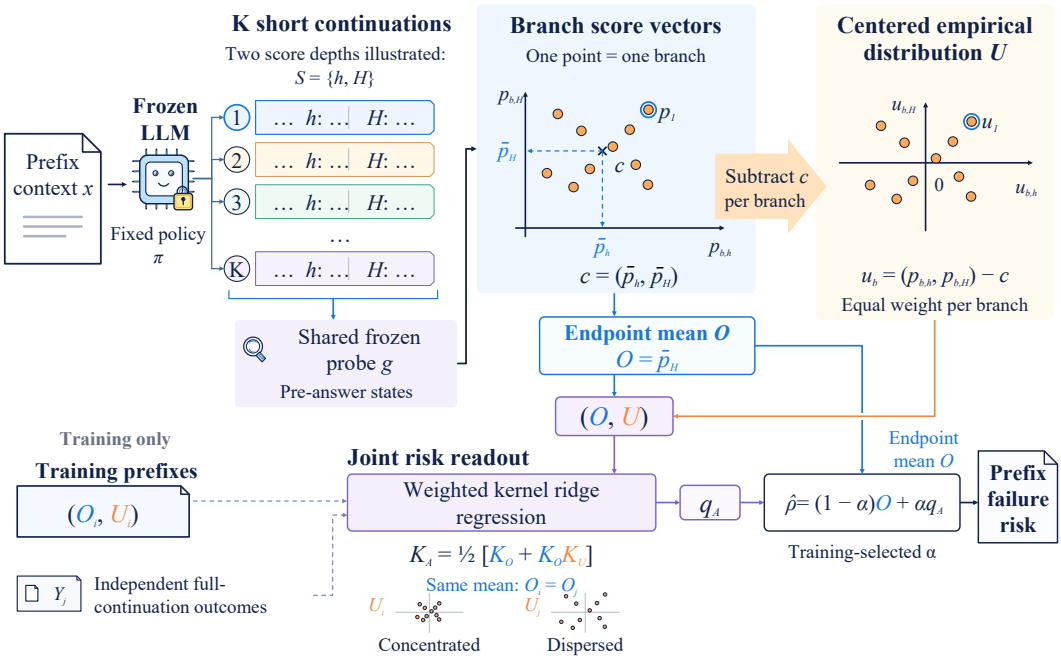

Figure 1: **Overview of TrajUQ.** A shared frozen probe scores K short continuations from prefix x under a fixed policy. The endpoint mean O and the equal-weight empirical distribution U of complete, coordinate-wise centered branch-score vectors feed a joint risk readout. Independently sampled full-continuation outcomes provide fitting targets (dashed paths). The fitted readout is combined with O using a training-selected coefficient α to predict prefix failure risk $\widehat{\rho}$. Two score depths, $S = \{h, H\}$, are illustrated. The concentrated and dispersed examples share the same mean, coordinate scales, and point count; their relationship to risk is learned from independent outcomes.

unfiltered continuations under the fixed policy, ideal conditional failure scores recover prefix risk in expectation (Setlur et al., 2025) (Appendix A.1). For finite probe scores, we use O as the level component of the learned risk readout.

To retain variation across branches, we collect each branch's scores at depths S and center each coordinate across branches:

$$u_b = (p_{b,t} - \bar{p}_t)_{t \in S}, \qquad b = 1, \ldots, K, \tag{2}$$

where $\bar{p}_t$ is the mean score at depth t. We assign equal mass to these K vectors in empirical distribution U. Grouping scores by branch preserves dispersion at each depth and joint variation across depths, while the distribution is invariant to branch ordering. Centering removes each depth's mean from U; the endpoint mean O supplies the level separately, and the other coordinate means are used only for centering.

With finite, imperfect probe scores, a mean-only readout cannot distinguish collections with the same O but different branch variation (Figure 1). We learn from independent continuation outcomes whether U helps distinguish their risks beyond O. Both components reuse the observed branches, motivating our test of better prediction with fewer branches.

4.2 A JOINT RISK READOUT

For this joint readout, we compare empirical distributions through kernel mean embeddings, following Szabó et al. (2016). We compute embedding inner products by averaging point-kernel similarities between complete centered score vectors over all cross-branch pairs. We normalize by the product of branch counts to compare different-sized collections, each centered on its own observed branches. For prefix records i and j, a Gaussian kernel K_U compares these embeddings, and a scalar Gaussian kernel K_O compares endpoint means through standardized $\mathrm{logit}(O)$. We combine them as

$$K_A(i, j) = \tfrac{1}{2} K_O(i, j)\big[1 + K_U(i, j)\big]. \tag{3}$$

We retain the additive mean-based K_O component, while $K_O K_U$ links distribution similarity to endpoint-mean similarity. The learned response to branch variation can therefore depend on score level. We determine coordinate scales and bandwidths from training data and reuse them for prediction. Appendix B gives the kernel calculations and numerical rules.

We fit the joint readout by weighted kernel ridge regression to the empirical failure rates r_i from independent label continuations, and clip its output to obtain $q_{A,i} \in [0, 1]$. Our final prediction is $\widehat{\rho}_i = (1 - \alpha)O_i + \alpha q_{A,i}$, where $\alpha \in [0, 1]$ controls the learned readout's contribution relative to the raw mean. We select α using nested out-of-fold predictions on training questions, minimizing the question-weighted squared error against r_i. Appendix B.4 specifies the regression weights and objective; Appendix B.5 gives the shrinkage rule.

4.3 Training and a matched mean control

We nest fitting within question-grouped splits. The coordinate mask remains fixed across folds. For each inner validation group, we fit the scaler, principal component analysis (PCA), probe, and regressor with that group excluded. We cross-fit regressor-training scores with probes that additionally exclude the questions supplying those scores (Appendix B.6). Probes and risk regressors are fitted and evaluated separately within each model–task pair. For each readout and outer fold, we share one shrinkage coefficient estimated from inner out-of-fold predictions pooled across all model–task pairs. Our nested-tuned comparison also shares the training-selected kernel candidate across pairs (Appendix B.7).

For matched mean readout M, we replace K_A with K_O, predicting $c(O)$ from the endpoint mean. We use the same scores, independent targets, question splits, and weighted fitting, clipping, and shrinkage rules for both readers, fitting their coefficients and shrinkage separately. The raw endpoint mean O remains in each final mixture. Our comparison evaluates the distribution input and joint kernel readout together. Section 5.4 further matches training-side parameter selection and compares alternative score representations and their induced readouts.

5 Experiments

We compare readouts using the same observations, then examine whether fewer observations support effective prediction and offline screening. At the reference budget, our matched comparisons examine score level and branch variation. Broader baselines, observation allocation, and representation sensitivity complement these results before an application to offline candidate selection.

5.1 Evaluation design

We evaluate 540 original questions: 180 each from MMLU-Pro (Wang et al., 2024b), StrategyQA (Geva et al., 2021), and TriviaQA (Joshi et al., 2017). Thirteen model configurations and five prefix sites per question yield 35,100 records. We place the sites at 10%, 25%, 50%, 75%, and 90% of each archived anchor's generated tokens; prefix inputs end before its `Final Answer:` marker. We use these fractions to organize offline evaluation; probe and readout inputs exclude absolute indices, fractions, and completed-trace lengths. Continuation states likewise precede their branch's marker. Appendix B.1 gives the collection policy, checkpoint revisions, and labels; Appendix B.2 specifies endpoint selection.

Our default observation uses $K = 10$, $H = 64$, and requested score depths $S = \{16, 64\}$. Our labels come from ten independent full continuations per prefix. Appendix C.3 separates outcome types and format compliance. Our five task-stratified outer folds keep each original question's records together. For each outer fold, we fit and select parameters using only the training questions, with validation within that partition (Appendix B.6).

We use two fitting protocols for different comparisons. Our *fixed reference* uses training-median bandwidths and ridge penalty $\lambda = 0.01$; its fitted probe, readouts, and shrinkage remain fixed throughout the budget scans. Our *nested-tuned comparison* selects among nine training-side candidates for each representation before refitting (Section 5.4). Both protocols are evaluated against separately sampled continuation outcomes using the original question weights; Appendix B specifies their fitting and shrinkage rules.

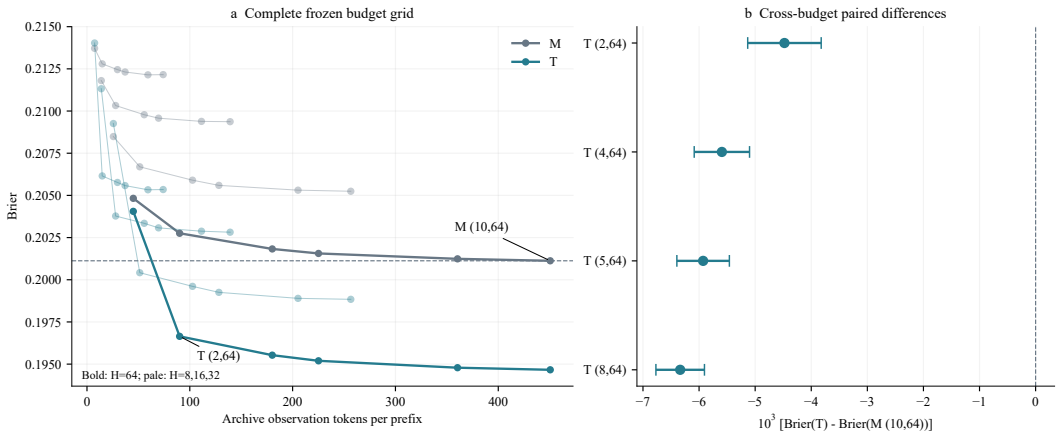

Figure 2: **Readout design and observation requirements in the frozen pipeline.** Left: all 24 allocations for the learned mean (M) and TrajUQ (T). Lines connect increasing branch counts at fixed depths; bold lines use $H = 64$, and pale lines $H \in \{8, 16, 32\}$. The dashed line is M(10,64). Right: T(K,64) minus M(10,64) Brier differences, multiplied by 10^3, for four lower-budget configurations. Bars retain the 95% simultaneous bootstrap intervals for these $K \geq 2$ contrasts under the original 24-contrast calibration (Appendix H). Each fitted predictor stays fixed throughout the scan and the 2,000 original-question bootstrap draws.

We evaluate Brier loss against individual Bernoulli labels with question-balanced aggregation as our primary metric. Negative log-likelihood (NLL) complements probability scoring, and area under the receiver operating characteristic curve (AUROC) measures ranking within model–task–prefix strata. Our paired percentile intervals use 2,000 task-stratified original-question bootstrap draws with predictions fixed. For TrajUQ (T), we use a 95% interval in the primary nested-tuned comparison with the learned mean (M), and a conservative 97.5% interval for comparison with the second-order summary readout (L). Our post-hoc cross-budget analysis reports the 20 $K \geq 2$ contrasts while retaining simultaneous intervals calibrated over the original 24-contrast family (Appendix H). Appendix C.2 defines the weights, finite-label scoring, and original interval conventions.

5.2 Risk prediction at matched observation budgets

We freeze the fixed-reference probe, readouts, and shrinkage fitted at $K = 10, H = 64$ while query observations vary across 24 combinations of branch count $K \in \{1, 2, 4, 5, 8, 10\}$ and requested depth $H \in \{8, 16, 32, 64\}$ tokens. Each matched comparison, denoted T(K, H) versus M(K, H), uses the same branches and depth. Width follows nested subsets in archived branch order; depth is capped by pre-answer length. We count observation tokens from saved continuations as $C_{\text{obs}} = \sum_b \min(H, G_b)$, where G_b is the recorded number of generated tokens on that branch (Appendix F.1). Training collection, state scoring, and readout computation are accounted for separately from these observation tokens (Appendix F.3).

Our T readout has lower Brier point estimates than M in 22 of the 24 allocations with matched observations (Figure 2, left; Table 16). Averaging equally over the 20 multi-branch allocations ($K \geq 2$), with the original question weights retained within each allocation, our mean Brier decreases from 0.2074 to 0.2009: an absolute reduction of 0.006482 and a mean per-allocation relative reduction of 3.125%. With both readers using two branches and $H = 64$, T–M is -0.006108, with ordinary 95% paired interval $[-0.006398, -0.005804]$.

At $K = 1$, the paired T–M intervals include zero; U is a point mass at zero, leaving the readers as separately fitted frozen mappings of the endpoint score (Appendix E.2).

5.3 Reducing observation requirements

We test whether lower-cost T can outperform M(10,64), keeping both reference fits frozen. In our post-hoc scan, seven configurations with at least two branches have simultaneous 95% Brier-difference

intervals below zero: $K \in \{2, 4, 5, 8\}$ at $H = 64$ and $K \in \{5, 8, 10\}$ at $H = 32$. Their token savings span 20.01%–80.00%, computed as $1 - C_{\text{obs}}(K, H)/C_{\text{obs}}(10, 64)$. These $K \geq 2$ intervals retain the original 24-contrast adjustment (Table 22).

For example, our two-branch T at $H = 64$ attains Brier 0.1966 with 90.10 observation tokens per prefix, compared with 0.2011 and 450.5 tokens for M(10,64): a 2.226% relative Brier reduction with approximately 80% fewer tokens. The paired difference is -0.004478, with simultaneous 95% interval $[-0.005133, -0.003823]$ (Figure 2; Appendix H).

For this comparison, all three task intervals are below zero. All thirteen model configurations have negative point differences; the descriptive interval for Qwen2.5-14B crosses zero (Table 23). Thus the pooled observation advantage extends across the tasks, while the model-specific evidence varies in strength.

Low-budget offline risk screening. Our offline screening tests whether the lower-budget reader selects prefixes with lower empirical failure. Using fixed-reference predictions, we rank prefixes within model–task–prefix strata in this post-hoc analysis, retaining equal prefix fractions (coverage) for T(2,64) and M(10,64). We estimate retained-set empirical failure rates from independent continuation outcomes; tied scores receive a common label-independent fractional inclusion (Appendix H.3). Complementary screening and outcome-wise loss analyses at the default budget use nested-tuned predictions (Appendix G; Figure 6).

At 50% coverage, our retained failure rates are 39.97% for two-branch T and 43.48% for ten-branch M. The paired difference is -3.517 percentage points, with descriptive 95% interval $[-3.918, -2.958]$. Figure 3 and Table 24 show the full coverage grid.

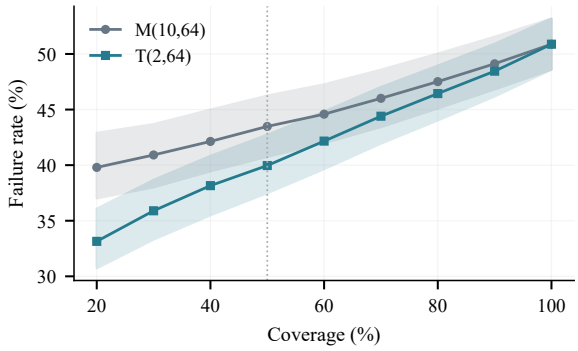

Figure 3: **Low-budget offline screening.** Frozen T(2,64) and M(10,64) retain equal coverage within each model–task–prefix stratum. Bands: descriptive pointwise 95% question-bootstrap intervals, with selection recomputed per draw. The dotted line marks 50% coverage; both curves meet at 100%. The generation policy is unchanged.

5.4 Readout design at the reference budget

We assess the joint design at $K = 10, H = 64$ against alternatives based on means, second-order summaries, or centered branch variation, using nested tuning. M learns from the endpoint mean O; the complementary post-hoc M2 adds the earlier mean m_{16}. L adds the variances v_{16}, v_{64} and covariance $v_{16,64}$, while TrajUQ (T) adds the empirical distribution U of the centered score vectors. Our L and M2 readouts retain T's mean-plus-product architecture with their respective summary kernels. For all readers, we use shared cross-fitted scores, independent targets, and the nine-candidate search. For post-hoc U, we use K_U and shrink toward a training-side constant, excluding O from both its kernel and final mixture; the other readers retain their O-based shrinkage (Appendix D.4). Appendix B.7 defines the protocol, and Appendix D.3 gives M2's definition and task results.

With T, we reduce Brier by 3.21% relative to M, with paired difference -0.006461 and 95% interval $[-0.006787, -0.006127]$. T also improves on L's tested second-order summary (Table 1). Within the declared search, we compare joint predictive designs that differ in both representation and induced kernel. Table 9 gives the corresponding task-level Brier, NLL, and AUROC; Appendix D.2 gives the complementary L–M contrast.

Adding the earlier mean in M2 leaves Brier close to M (0.201). Our comparison with M2 favors T, which uses the early mean only for centering (Table 1; Appendix D.3).

Table 1: Nested-tuned readouts at $K = 10, H = 64$ with the same nine-candidate search. Upper: Brier $\pm$ SE. Lower: 10^3 times T-minus-control differences and paired intervals. T–M2 and T–U are descriptive post-hoc comparisons (Appendices D.3 and D.4).

Readout	Input representation	Pooled Brier
M	O	0.2011 ± 0.003158
U	U	0.2119 ± 0.003147
M2	O, m_{16}	0.2010 ± 0.003155
L	$O, v_{16}, v_{64}, v_{16,64}$	0.2009 ± 0.003149
T	O, U	0.1947 ± 0.003039
Contrast	$10^3\Delta$	Paired interval
T–M	-6.461	[-6.787, -6.127] (95%)
T–M2	-6.383	[-6.685, -6.042] (95%)
T–L	-6.249	[-6.527, -5.947] (97.5%)
T–U	-17.27	[-19.11, -15.58] (95%)

Removing score level gives the U control, with Brier 0.2119. Our T–U comparison, together with T–M, supports retaining both score level and centered branch variation in the tested readout (Table 1).

Fixed-reference comparison. In Table 2, we compare the reference fit used for the budget scans with broader baselines (Appendix D.1). The mean control is learned from O; other methods retain their own observations, mappings, and acquisition requirements (Appendix C).

Table 2: Fixed-reference probability prediction. Brier $\pm$ question-bootstrap SE on 540 questions. The last two methods share $K = 10, H = 64$ observations and fixed fitting rules; other methods retain their own inputs and mappings. Bold marks column minima.

Method / Dataset	MMLU-Pro	StrategyQA	TriviaQA	Pooled
LN-NLL	0.2498 ± 0.0006	0.2544 ± 0.0009	0.2442 ± 0.0006	0.2495 ± 0.0004
EigenScore	0.2425 ± 0.0013	0.2545 ± 0.0012	0.2384 ± 0.0017	0.2451 ± 0.0008
SEP-style Probe	0.2545 ± 0.0011	0.2549 ± 0.0008	0.2406 ± 0.0006	0.2500 ± 0.0005
Effective Rank	0.2417 ± 0.0014	0.2532 ± 0.0012	0.2389 ± 0.0018	0.2446 ± 0.0009
Sample Consistency	0.2459 ± 0.0027	0.2335 ± 0.0033	0.2049 ± 0.0017	0.2281 ± 0.0015
Semantic Entropy	0.2639 ± 0.0029	0.2337 ± 0.0020	0.1996 ± 0.0020	0.2324 ± 0.0013
KLE	0.2616 ± 0.0022	0.2448 ± 0.0015	0.2105 ± 0.0016	0.2390 ± 0.0010
Semantic Density	0.2462 ± 0.0026	0.2364 ± 0.0030	0.2069 ± 0.0017	0.2298 ± 0.0014
SRE-UQ	0.2536 ± 0.0011	0.2451 ± 0.0009	0.2269 ± 0.0011	0.2419 ± 0.0006
Current-prefix Probe	0.2254 ± 0.0050	0.2175 ± 0.0077	0.2036 ± 0.0030	0.2155 ± 0.0033
Matched mean $c(O)$	0.2122 ± 0.0047	0.2073 ± 0.0075	0.1838 ± 0.0028	0.2011 ± 0.0032
TrajUQ	$\mathbf{0.2055} \pm 0.0045$	$\mathbf{0.2014} \pm 0.0072$	$\mathbf{0.1771} \pm 0.0028$	$\mathbf{0.1947} \pm 0.0030$

Our Current-prefix Probe uses its own fitted mapping of the current prefix state, yielding pooled Brier 0.2155. At $K = 10, H = 64$, both short-continuation readers have lower Brier: 0.2011 for $c(O)$ and 0.1947 for TrajUQ. Compared with the matched learned mean, we reduce Brier with TrajUQ by 3.21%, with paired difference -0.006461 and 95% interval $[-0.006759, -0.006150]$.

Our Brier reductions are 3.20% on MMLU-Pro, 2.85% on StrategyQA, and 3.63% on TriviaQA relative to the matched mean readout. All thirteen model configurations have negative paired Brier intervals (Table 8). Our complementary metrics show better probability scoring and within-stratum risk ranking: pooled NLL decreases by 2.65%, and AUROC increases from 0.6160 to 0.6851. Appendix D.1 gives their paired intervals.

5.5 Observation allocation and representation sensitivity

Within the frozen pipeline, we also compare how T allocates its observations between branch count and depth. At the common nominal ceiling $KH = 320$, we obtain 1.83% lower Brier with T(5,64) than with T(10,32), using 225.2 versus 256.6 observation tokens, a 12.24% reduction. At fixed

$H = 64$, T(8,64) reduces tokens by 20.01% relative to T(10,64), trading a small Brier increase for lower observation cost (Table 18).

Full depth and width scans (Figure 5) report paired effects for both readers (Appendix E.2). Our position-subset refits show that retained depths matter: a three-position subset has lower pooled loss than all four positions (Appendix E.1). With upstream probes refitted and the downstream readers frozen, we observe T's advantage at all seven tested probe inputs (Table 15).

5.6 Reducing Answer Failure from Limited Observations

To test whether limited observations support more reliable answer selection, we adapt the frozen TrajUQ readout in a post-hoc TriviaQA analysis to rank ten saved continuations from a common prefix, using each candidate's probe score and the shared pool distribution. We use probe scores from pre-answer states within the first 64 generated tokens of each candidate; we return the lowest-scoring candidate's saved answer at the final checkpoint. We evaluate Single answer by the expected failure of uniform selection from this pool. Across 180 questions and 13 model configurations, we reduce failure with TrajUQ from 56.47% for Single answer to 51.92% (paired 95% interval: $[-5.547, -3.568]$ percentage points). Direct branch-probe selection yields 53.16% failure; the paired 95% interval for TrajUQ minus Branch probe is $[-1.853, -0.819]$ percentage points. Current-prefix Probe applied to candidate states yields 55.47% failure, while GUARD-style selection by mean entropy over full saved continuations yields 54.36% (Figure 4; Appendix I).

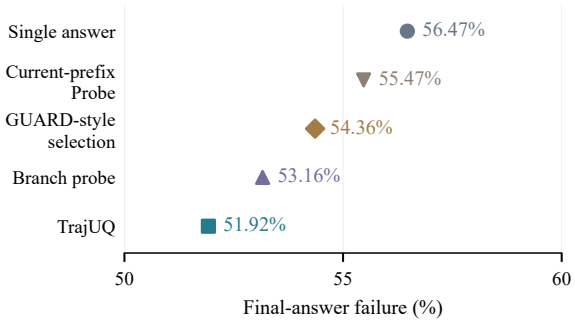

Figure 4: **TriviaQA candidate selection.** Final-answer failure across 180 questions and 13 model configurations. Single answer samples uniformly; Branch probe ranks branch scores directly, and Current-prefix Probe uses its separately fitted predictor. GUARD-style selection applies mean-entropy ranking to full continuations; TrajUQ uses the 64-token checkpoint, capped at the pre-answer boundary. Paired intervals appear in Appendix I.

6 Discussion

With TrajUQ, we combine endpoint means and centered branch-score distributions to predict prefix failure risk. Our matched comparisons support the representation and induced kernel as a joint design that improves prediction from the same observations, while our frozen-budget scans show that better readout can reduce observation requirements. Our outcome breakdown shows lower prediction losses for both correct and incorrect parsed answers (Table 20; Appendix G). Our offline screening selects lower-risk prefixes, while TriviaQA candidate selection improves final-answer accuracy without updating LLM parameters.

For models with known architectures and accessible hidden representations, we can fit the probe and risk readout to a target task using labeled continuation outcomes. Our evaluation covers 540 questions under the stated tasks, model configurations, and generation policies. We fit and evaluate predictors within each model–task configuration, with pooled scores summarizing those evaluations. This dataset informed our method development; within the reported evaluation, parameter fitting and selection exclude the corresponding outer evaluation fold. Evaluation on new datasets, model configurations, and sampling policies would clarify the breadth of these gains.

Our budget and screening analyses keep predictors fixed after fitting at the reference budget, with a fixed branch order and cost measured in observation tokens. We could next adapt observation depth and branch count to each prefix. Jointly evaluating prediction quality, generation time, state access, and readout computation would help choose observation budgets for individual prefixes.

Together, these results link effective risk readout to fewer observations and more reliable answer selection.

AI USE STATEMENT

We used generative AI tools to aid manuscript preparation and polish the writing, including translating and editing text, improving clarity and readability, organizing sections and tables, and assisting with reference formatting and LaTeX editing.

The authors reviewed and revised all AI-assisted edits to ensure that they accurately reflected the underlying research, mathematical arguments, implementation, and experimental results. The authors take full responsibility for the final content, claims, and artifacts in this work.

ETHICS STATEMENT

This study uses existing question-answering benchmarks and pretrained language models for methodological evaluation of prefix-risk prediction. It involves no new collection of personal data or direct interaction with human participants.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the target, supervision, and readout; Appendix A gives notation and the ideal mean interpretation. Appendix B specifies collection and fitting, including the task instructions (Appendix B.1) and answer-boundary state indices (Appendix B.2). Appendix C supplies baseline and scoring protocols; Appendices D, E, and F provide full metrics, ablations, and cost results. The supplementary package provides selected core routines and machine-readable aggregate results. The complete implementation and supporting experimental artifacts will be released upon acceptance.

REFERENCES

Mohit Singh Chauhan, Vipin Gyanchandani, and Dylan Bouchard. When do supervised UQ ensembles improve LLM hallucination detection? A robustness study. arXiv preprint arXiv:2608.24492, 2026. URL https://arxiv.org/abs/2608.24492.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations*, pp. 34979–35010, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9682490bedc064aba8aac1ab3f703507-Abstract-Conference.html.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Nils Grünefeld, Bertram Højer, Philipp Mondorf, Barbara Plank, Anna Rogers, Christian Hardmeier, Stefan Heinrich, and Jes Frellsen. Tracing uncertainty in language model "reasoning". arXiv preprint arXiv:2605.07776, 2026. URL https://arxiv.org/abs/2605.07776.

Xinzhe Huang, Biwu Yao, Kedong Xiu, Mengnan Zhao, Di Wang, Puning Zhao, and Tianhang Zheng. ProbGuard: Calibrated safety risk estimation from LLM output distributions. arXiv preprint arXiv:2608.10621, 2026. URL https://arxiv.org/abs/2608.10621.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. arXiv preprint arXiv:2406.15927, 2024. URL https://arxiv.org/abs/2406.15927.

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.

Jinyuan Li, Langlin Huang, Chengsong Huang, Shaoyang Xu, Donghong Cai, Yuyi Yang, Wenxuan Zhang, and Jiaxin Huang. Process rewards with learned reliability. arXiv preprint arXiv:2605.15529, 2026. URL https://arxiv.org/abs/2605.15529.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 19260–19268, 2025. doi: 10.1609/aaai.v39i18.34120. URL https://ojs.aaai.org/index.php/AAAI/article/view/34120.

Rachel Ma, Dylan Hadfield-Menell, and Kristjan Greenewald. Distributional process reward models: Calibrated prediction of future rewards via conditional optimal transport, 2026. URL https://arxiv.org/abs/2605.06785.

Uzay Macar, Paul Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting LLM reasoning requires resampling. In *International Conference on Learning Representations*, pp. 92567–92600, 2026. URL https://openreview.net/forum?id=bVsAuIOvJ5.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929, 2024. doi: 10.52202/079017-0283. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/10c456d2160517581a234dfde15a7505-Abstract-Conference.html.

Young-Jin Park, Kristjan Greenewald, Kaveh Alimohammadi, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38852–38895, 2025. doi: 10.52202/085713-1297. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/37664246a1e07e212ddacea6e5a523f2-Abstract-Conference.html.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, pp. 134507–134533, 2024. doi: 10.52202/079017-4274. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f26d4fbaf7dfa115f1d4b3f104e26bce-Abstract-Conference.html.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, pp. 60808–60838, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/98711dea460bdefe0e651ca23ec98ba2-Paper-Conference.pdf.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40088–40106, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1862. URL https://aclanthology.org/2026.acl-long.1862/.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. URL https://jmlr.org/papers/v17/14-510.html.

Pragatheeswaran Vipulanandan, Kamal Premaratne, and Dilip Sarkar. Semantic uncertainty quantification of hallucinations in LLMs: A quantum tensor network based method. In *International Conference on Learning Representations*, pp. 46061–46100, 2026. URL https://openreview.net/pdf/d00c6f1f6d607af7a3e0d0bfeebefe57d49fa18f.pdf.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Rui Wang, Zeming Wei, Guanzhang Yue, and Meng Sun. Revisiting hallucination detection with effective rank-based uncertainty. arXiv preprint arXiv:2510.08389, 2025. URL https://arxiv.org/abs/2510.08389.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024b. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Yuhang Zhou, Yixin Cao, and Guangnan Ye. From correctness to utility: Gain-based prefix evaluation for LLM reasoning. arXiv preprint arXiv:2606.07190, 2026. URL https://arxiv.org/abs/2606.07190.

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

A Notation and Supervision

For record i, the single-prefix symbols n, Y_j, and r become n_i, Y'_{ij}, and r_i, respectively. Likewise, the branch states $Z_{b,t}$ and scores $p_{b,t}$ are written as $Z_{ib,t}$ and $p_{ib,t}$ when the record index is included. The endpoint mean is denoted by O for a single prefix and O_i for record i.

Table 3: Notation. K counts feature branches and n_i counts independent label continuations. Neither is the number of independent original questions.

Symbol	Meaning
x, x_i, π	Question and reasoning prefix (record i when indexed), and fixed model and decoding policy
$F_{A,i}, L_i$	Feature branch pool and independent label pool
$Y, Y_{ib}^{\text{own}}, Y'_{ij}$	Generic failure indicator, observed feature-branch outcome, and independent label-continuation outcome
$\rho(x), \rho_i, r_i, C_i$	Prefix failure risk, its value $\rho_i = \rho(x_i)$ for record i, observed failure rate, and failure count
H, K	Observation horizon and number of feature branches
S, h	Selected score depths for U, and earlier depth in the two-position configuration
$g, p_{b,t}, p_{ib,t}$	Frozen branch probe and its scalar score (record i when indexed) at requested depth t
O, O_i	Endpoint average score (record i when indexed)
u_b, u_{ib}, U, U_i	Centered score vector and its empirical distribution (record i when indexed)
K_O, K_U, K_A	Endpoint-mean, distribution, and joint kernels
v_i, w_i	Question-balancing weight and normalized KRR weight proportional to $v_i n_i$
$q_A, \widehat{\rho}, \alpha$	Clipped joint prediction, final TrajUQ prediction, and OOF mixing coefficient
$c(O_i)$	Mean-based prediction after matched scalar KRR and OOF calibration

We use feature-branch outcomes to supervise the probe, and independent label outcomes to supervise and score the risk readout. The prediction inputs consist of feature-branch scores, from which U_i is constructed.

A.1 Ideal mean-score interpretation

The endpoint mean is $O = \bar{p}_H = K^{-1} \sum_{b=1}^{K} p_{b,H}$. Math-Shepherd's soft estimator averages success indicators of full continuations (Wang et al., 2024a). For short observations, the policy-value view of future correctness (Setlur et al., 2025) gives an analogous interpretation using ideal conditional failure scores. Consider a continuation sampled from x under π without filtering on its eventual outcome. Let Z_H and Y denote its observed endpoint state and final failure indicator, respectively. By the law of total expectation,

$$\mathbb{E}_{Z_H|x,\pi}\Big[\Pr_{\pi}(Y = 1 \mid x, Z_H)\Big] = \Pr_{\pi}(Y = 1 \mid x) = \rho(x). \tag{4}$$

Thus ideal conditional failure scores recover prefix risk in expectation. In practice, probe scores can differ from these conditional probabilities, and O averages finitely many branches. Our empirical comparison therefore tests the learned joint readout against a matched mean readout under the actual observation and scoring procedure.

B Implementation and Training Details

B.1 Collection policy, questions, and outcome labels

The thirteen checkpoints and their revisions are listed in Table 4; we load each tokenizer from the same checkpoint revision. For generation, we use temperature 0.7, top-p 0.95, at least one new token, and maximum lengths of 256, 192, and 128 tokens for MMLU-Pro, StrategyQA, and TriviaQA, respectively. Full continuations stop at the model's EOS rule or the length cap.

For Qwen2.5, we use top-$k = 20$ and repetition penalty 1.05. Llama-3.1 and Yi-1.5 use top-$k = 50$ and repetition penalty 1.0, based on their generation files and versioned library defaults. For Gemma3, we use top-$k = 64$, repetition penalty 1.0, and EOS IDs 1 and 106; its padding ID is the tokenizer's EOS ID, 1. Our collection records retain the EOS and padding rules and the source of each field, distinguishing explicit collector settings, checkpoint metadata, and versioned library defaults.

Our prompts contain raw task instructions followed by the question, with lettered options for MMLU-Pro; we apply no chat template. The exact task instructions appear below.

Task prompt instructions.

> **MMLU-Pro.** `Answer the multiple-choice question using concise reasoning. The last line must be exactly: Final Answer: <A-J>`
>
> **StrategyQA.** `Answer the yes/no question using concise multi-hop reasoning. The last line must be exactly: Final Answer: <yes/no>`
>
> **TriviaQA.** `Answer the trivia question using concise reasoning. Give only a short entity or phrase on the final line. The last line must be exactly: Final Answer: <short answer>`

Each anchor and probe-input branch in this evaluation contains a single literal `Final Answer:` marker following a nonempty reasoning segment. Our input reader separates this segment from the marker and final-answer tokens before prefix construction and continuation-state selection. We confine marker search to the generated portion, excluding task-prompt tokens. We sample continuations directly from the recorded tokenized prefix, which already contains the task prompt. State extraction receives the nonempty pre-answer token segment and applies the horizon rule in Appendix B.2. To label outcomes, we parse the complete continuation separately, using answer-extraction rules distinct from the feature-input boundary.

For MMLU-Pro, we sample across subject categories (seed 20260625). For StrategyQA, we balance yes/no labels within source splits (seed 20260707). For TriviaQA, we deduplicate question IDs, remove train IDs from the validation pool, shuffle the two pools (seed 20260717), and retain answer aliases. The 110/35/35 counts refer to collection-time partitions, distinct from the upstream source splits. Our evaluation determines training, validation, and test membership through the question-grouped outer and inner folds over all 180 questions per task (Appendix B.6).

Our independently generated anchor response defines five prefix sites at fractions 0.10, 0.25, 0.50, 0.75, and 0.90 of its generated token sequence, measured without the task-prompt tokens. The recorded fractional sites determine the prefixes, whose usable inputs stop before the anchor's final-answer marker. From each prefix, we sample twenty feature continuations and ten independent label continuations. We record four requested horizons, $\{8, 16, 32, 64\}$; these observation depths are distinct from the default readout's score set $S = \{16, 64\}$. Our fixed partition assigns ten feature branches to F_A and ten completed responses to F_B, with F_A supplying the default $K = 10$ observations. Seeds hash the base seed, task, question, prefix, pool name, and branch index, giving the feature and label streams distinct keys. Our collector fills fixed branch-ID slots and resumes from saved records; correctness is evaluated after generation and does not trigger branch replacement.

Our evaluator uses the last valid `Final Answer` or `Correct Answer` declaration, falling back to the terminal line. MMLU-Pro compares normalized option letters; StrategyQA compares yes/no, mapping true/false to yes/no; TriviaQA lowercases, removes punctuation and articles, collapses whitespace, and checks exact membership in the stored alias set. We count a response as a failure when the evaluator extracts no normalized answer. Format validity is recorded separately. Length-truncated responses use the same extraction and correctness rules.

B.2 Frozen branch probe

Our probe pipeline applies a fixed finite-coordinate mask, StandardScaler, 32-component whitened PCA, and logistic regression to training-branch failure labels with branch weights. Logistic regression uses $C = 1$, L-BFGS, tolerance 10^{-4}, and at most 1,000 iterations. Randomized PCA uses seed 0, ten oversamples, four power iterations, and QR normalization. Our scaler, PCA, and probe fits use the permitted training questions in each scope. For record i, each of its ten training branches has weight $v_i/10 = 1/650$, using the question-balancing record weight $v_i = 1/65$. We pass these weights directly to logistic regression as sample weights, without rescaling; StandardScaler and PCA are fitted without sample weights. Our H64-trained pipeline scores both H16 and H64 states. The last stored slot is zero-based index 3; Table 4 gives its absolute transformer-layer mapping.

Table 4: Checkpoints and final-layer mapping from collection manifests. The final stored slot is 3 (zero-based); layer numbers index transformer hidden-state outputs. Full revision identifiers are retained in the collection records.

Checkpoint	Revision (prefix)	Layer
google/gemma-3-12b-it	96b6f1eccf38	48
google/gemma-3-27b-it	005ad3404e59	62
google/gemma-3-4b-it	093f9f388b31	34
meta-llama/Llama-3.1-70B-Instruct	1605565b47bb	80
meta-llama/Llama-3.1-8B-Instruct	0e9e39f249a1	32
Qwen/Qwen2.5-14B-Instruct	cf98f3b3bbb4	48
Qwen/Qwen2.5-32B-Instruct	5ede1c97bbab	64
Qwen/Qwen2.5-3B-Instruct	aa8e72537993	36
Qwen/Qwen2.5-72B-Instruct	495f39366efe	80
Qwen/Qwen2.5-7B-Instruct	a09a35458c70	28
01-ai/Yi-1.5-34B-Chat	fa4ffba162f2	60
01-ai/Yi-1.5-6B-Chat	771924d1c83d	32
01-ai/Yi-1.5-9B-Chat	1a0fc698cf88	48

Our coordinate mask excludes coordinates found nonfinite in the archive's original training feature grids. We define it without labels and reuse it without reselection across grouped folds, all four horizons, and branch counts. Scaler and PCA parameters are fitted within each permitted question scope. For Gemma3-4B, we exclude zero-based coordinate 443 from every row at both observation depths and for both readouts, leaving 2,559 of 2,560 coordinates in each task. All other model–task masks retain every coordinate. All rows are retained.

Answer boundary and state indices. For branch b, let $L_b \geq 1$ be the number of generated tokens in its separated pre-answer reasoning segment. For requested depth h, we use the observed length $\ell_{b,h} = \min(h, L_b)$. We take the state after its last observed token: continuation-relative index $\ell_{b,h} - 1$, or absolute index $P_b + \ell_{b,h} - 1$ in the concatenated prefix–continuation sequence, where P_b is the prefix token count and both indices are zero-based. The probe reads the stored layer specified in Table 4.

At observation horizon H, we use the state at length $\min(H, L_b)$ for the endpoint mean. Our depth sweep uses the earlier coordinate at length $\min(16, H, L_b)$; both score coordinates are retained when they select the same state. These lengths are separate from the raw generated-length field used for observation-cost accounting (Appendix F.1). We use complete continuation text for branch-outcome supervision and independent risk labels, while the risk-readout inputs are the pre-marker scores.

B.3 Training-side coordinate scales and bandwidths

For coordinate a of the centered score vector, let m_a and s_a be the mean and standard deviation over the pooled training branch points. We define $z_a(u) = (u_a - m_a)/s_a$, replacing a standard deviation no larger than 10^{-12} by one. We use the point kernel

$$k_{\mathrm{pt}}(u, v) = \exp\left(-\frac{\|z(u) - z(v)\|^2}{2\tau_{\mathrm{pt}}^2}\right), \qquad \tau_{\mathrm{pt}}^2 = \mathrm{median}_+\{\|z(u) - z(v)\|^2\}_{\mathrm{train}}. \tag{5}$$

For records with K branches, the empirical embedding inner product, squared embedding distance, and Gaussian distribution kernel are

$$B_{ij} = \frac{1}{K^2}\sum_{b=1}^{K}\sum_{c=1}^{K} k_{\mathrm{pt}}(u_{ib}, u_{jc}),$$

$$D_{ij} = B_{ii} + B_{jj} - 2B_{ij}, \qquad K_U(i, j) = \exp\left(-\frac{D_{ij}}{2\tau_U^2}\right). \tag{6}$$

Here B_{ij} averages point-kernel similarities over branch pairs, with each complete centered score vector supplied to the point kernel. The distance D_{ij} compares the empirical kernel mean embeddings of the two centered score distributions.

We compute the positive median over distances above 10^{-12}. If none remain, we fall back to one. The distribution-kernel scale τ_U^2 is the positive median of the off-diagonal training distances D_{ij} in Eq. (6), using the same fallback. We check that distances exceed -10^{-10} before clipping negative roundoff to zero; a materially negative value indicates an implementation error.

For the endpoint-mean kernel K_O, we first clip O to $[10^{-6}, 1-10^{-6}]$, take its logit, and standardize using training-prefix inputs. A scalar Gaussian kernel then uses the positive median of squared training distances. For test inputs, we reuse every fitted transform and bandwidth. We fix the kernel normalization in Eq. (3) together with the ridge parameter to preserve regularization strength across comparisons.

For width sensitivity, we freeze the ten-branch training set while query bags may contain a different number of branches. The empirical inner product in Eq. (6) then uses $1/(K_iK_j)$ and sums over the K_i and K_j available points, respectively; the displayed expression gives the equal-size case. We compute each query mean and its centering from only the observed query branches. At $K_i = 1$, the centered query distribution is a point mass at zero for every prefix.

Our width sensitivity panel uses one nested sequence per prefix, taking the first K positions in the archived ten-branch ordering after string-sorting the retained rollout identifiers. Array positions are distinct from the original rollout identifiers. Our predictors and paired depths share this fixed ordering.

B.4 Weighted KRR with a free intercept

We fit the joint readout directly to the independent failure rates r_i. Let v_i be the question-balancing weight and n_i the number of label continuations for record i. We normalize the training weights as $w_i = v_i n_i / \sum_j v_j n_j$. We minimize the weighted kernel ridge regression objective

$$(\widehat{a}, \widehat{f}) = \arg \min_{a,\, f \in \mathcal{H}_{K_A}} \sum_i w_i \big(a + f(O_i, U_i) - r_i\big)^2 + \lambda \|f\|^2_{\mathcal{H}_{K_A}}, \quad \lambda > 0. \tag{7}$$

Here $\mathcal{H}_{K_A}$ is the function space induced by K_A, and the intercept a is unpenalized. We clip the fitted output to obtain $q_{A,i} = \mathrm{clip}_{[0,1]}(\widehat{a} + \widehat{f}(O_i, U_i))$ before shrinkage. Our mean readout uses the corresponding objective with K_O.

We solve the regularized linear system with a free intercept by Cholesky factorization. Our fixed-reference protocol uses $\lambda = 0.01$ and training-median bandwidths; the nested-tuned comparison selects among the candidates specified below.

Our matched mean control reads only O_i, using K_O and the same objective, question exclusions, clipping, and independently fitted OOF shrinkage. Its final probability is $c(O_i)$.

B.5 Out-of-fold shrinkage

For training-side OOF predictions, write $d_i = q_{A,i}^{\mathrm{OOF}} - O_i^{\mathrm{OOF}}$. Minimizing the weighted squared risk error along the segment between these two predictions gives

$$\widehat{\alpha} = \mathrm{clip}_{[0,1]} \left(\frac{\sum_i v_i d_i (r_i - O_i^{\mathrm{OOF}})}{\sum_i v_i d_i^2} \right). \tag{8}$$

If the denominator is zero, we set $\widehat{\alpha} = 0$. For each arm, we estimate one coefficient per outer fold from the pooled inner-OOF predictions across all 39 model–task nodes. This coefficient is shared across nodes, while their probes and KRR parameters are fitted separately. The final prediction is $\widehat{\rho}_i = (1-\alpha)O_i + \alpha q_{A,i}$.

B.6 Question-level nested fitting

With the coordinate mask held fixed, we partition each model–task node's 180 original questions into 144 for outer-fold training and 36 for evaluation. Our scaler, PCA, probe, mapper, and shrinkage fits use the training partition, which is further divided into five inner groups. All model instances and prefix positions of the same original question preserve these assignments.

For an inner validation group V, we fit its probe on the outer training questions excluding V. For a mapper-training group J within that inner fold, we produce its scores with a probe that excludes both V and J. Our archived construction uses one outer scope, five single-exclusion scopes, and ten pair-exclusion scopes per outer fold and model–task node. The outer-scope probe supplies inputs for the evaluation partition; the final mapper uses cross-fitted outer-training inputs and their independent risk labels.

B.7 Nested-tuned information comparison

Our nested-tuned comparison uses the archived cross-fitted scores at $K = 10$, $S = \{16, 64\}$, with a separate fit for each readout. M uses O. L uses O and the vector $(v_{16}, v_{64}, v_{16,64})$, where $v_t = K^{-1}\sum_b (p_{b,t} - \bar{p}_t)^2$ and $v_{16,64} = K^{-1}\sum_b (p_{b,16} - \bar{p}_{16})(p_{b,64} - \bar{p}_{64})$. We use $1/K$ normalization for these empirical moments. For L, we replace K_U by a Gaussian kernel on the training-standardized three-vector, retaining the same $\frac{1}{2}K_O(1 + K_U)$ architecture. TrajUQ (T) uses the complete centered score-vector distribution.

For every arm, we search a grid fixed before fitting, with nine candidates: bandwidth-squared multipliers $\beta \in \{1, 0.5, 2\}$ and $\lambda \in \{0.01, 0.1, 0.001\}$. We apply the multiplier to both the scalar mean kernel and the outer distribution/statistic kernel; T retains the training-median point kernel. For each candidate, Eq. (8) gives the optimal α on pooled inner-OOF scores. We select one candidate and coefficient per arm and outer fold and share them across the 39 nodes; node-specific coefficients and transforms are refitted on permitted outer-training inputs. Selection minimizes the pooled inner-OOF question-weighted squared-rate error. For ties within 10^{-12}, we prefer the original candidate $(1, 0.01)$, then larger regularization, bandwidth nearest one, then smaller bandwidth.

Table 5 gives the selected values. The parameter α weights the fitted readout relative to the raw mean in the final prediction. Our fixed-reference protocol uses $\beta = 1, \lambda = 0.01$ throughout, with $\alpha_M = (0.2862, 0.2983, 0.3044, 0.3613, 0.3154)$ and $\alpha_T = (0.3966, 0.3777, 0.3901, 0.3945, 0.3701)$ in outer folds 0–4.

Table 5: Selected bandwidth-squared multiplier, regularization, and shrinkage for each readout and outer fold. Each row is shared across model–task nodes; their fitted coefficients remain separate.

Outer fold	Readout	β	λ	α
0	M	2	0.01	0.3379
0	L	2	0.01	0.4055
0	T	2	0.01	0.4485
1	M	2	0.001	0.3091
1	L	2	0.01	0.3513
1	T	2	0.001	0.3188
2	M	2	0.01	0.3606
2	L	2	0.01	0.4099
2	T	2	0.01	0.4450
3	M	2	0.001	0.3616
3	L	2	0.01	0.4265
3	T	2	0.01	0.4573
4	M	2	0.01	0.3871
4	L	2	0.01	0.3902
4	T	2	0.01	0.4499

C Baselines and Evaluation Protocol

C.1 Baseline observations and fitting

Likelihood and internal-state measures. For LN-NLL, we average the length-normalized negative log-likelihood of ten branches observed at $H = 64$. For EigenScore, we use the mean log-spectrum of a regularized branch-representation covariance at $H = 64$, following INSIDE (Chen et al., 2024). Effective Rank uses the entropy of their normalized singular values (Wang et al., 2025); the implementation returns this entropy without exponentiating it. Our adapted SEP-style Probe

uses $H = 16$ representations with a training-only binary partition of sampled semantic entropy as supervision (Kossen et al., 2024). These methods use feature subpool F_A.

Current-prefix Probe. Our baseline predicts prefix-conditional failure risk directly from the hidden state of the current target prefix, without generating additional continuation branches. We fit a model–task-specific StandardScaler–PCA32–logistic model to the independent target-prefix label counts through a weighted binomial expansion, using outer-training questions. It provides a reference for assessing the predictive value of acquiring short-continuation observations.

Sampled-output measures. For Sample Consistency, we use the largest normalized-answer category frequency (Lyu et al., 2025). For Semantic Entropy, we aggregate bidirectional-entailment clusters (Kuhn et al., 2023); KLE uses a weighted semantic graph and a heat kernel with fixed $t = 0.3$ (Nikitin et al., 2024). Semantic Density evaluates similarity-weighted density around its declared reference candidate (Qiu & Miikkulainen, 2024). SRE-UQ uses the adjusted semantic Rényi-2 statistic and its frozen parameter-selection protocol (Vipulanandan et al., 2026). We compute these statistics from ten completed continuations in feature subpool F_B, distinct from the independent ten-continuation label pool. For Semantic Entropy, KLE, and Semantic Density, we share a frozen bidirectional NLI evaluator.

Numerical baseline conventions. Our baseline implementations adapt the cited methods to the shared archive. EigenScore regularizes covariance by $10^{-3}I$; Effective Rank uses a singular-value floor of 10^{-8}. Our SEP-style binary target minimizes within-class squared error of training semantic entropy. Semantic Entropy weights samples by normalized $\exp(-\text{average NLL})$. For Semantic Density, we use the minimum-average-NLL response as reference (branch-ID tie break) and unique F_B responses as its candidate set. We compute its directional similarity as $1 - p_C - \frac{1}{2}p_N$, averaged over directions, where p_C and p_N are the NLI probabilities of contradiction and neutrality. Our shared NLI model is `microsoft/deberta-v2-xlarge-mnli`, revision `2b65bccdedce`, with maximum input length 512 and a slow DeBERTa tokenizer. Labels are contradiction/neutral/entailment in indices 0/1/2; semantic clustering requires entailment in both directions. For SRE-UQ, we use the published full Bernoulli-KL/inverse-UQ adjustment, selecting its $\{5, 15, 25, 50, 75, 100\}$ penalty grid by inner-OOF AUROC; its frozen backend uses seed 42 and 200 perturbations.

Fitting and comparison scope. We score all methods against the same target-prefix label outcomes with common question weights. Our nine uncertainty baselines retain their native-score ranking fields. We select their probability calibrator among identity, temperature, and positive-slope Platt maps using inner-OOF NLL; candidates within 10^{-4} of its minimum are ordered by Brier, complexity, then name. The current-prefix probe uses its direct probability for both scoring and ranking. Our broader comparison is descriptive because inputs, fitted mappings, and acquisition costs differ; $c(O)$ and TrajUQform the matched short-continuation comparison.

C.2 Probability scoring and aggregation

Let C_i count failures among the n_i independent label continuations for record i, with $r_i = C_i/n_i$. The mean Bernoulli Brier for these trials is

$$\frac{1}{n_i}\sum_{j=1}^{n_i}(q_i - Y'_{ij})^2 = q_i^2 - 2q_i r_i + r_i = (q_i - r_i)^2 + r_i(1 - r_i). \tag{9}$$

The last term is shared by methods scored on the same labels and weights. Paired Brier and squared-rate-error differences are therefore equal, although the absolute values of the two losses differ. Conditional on the prefix and a prediction formed without label-pool access, averaging over independent label continuations yields $(q_i - \rho_i)^2 + \rho_i(1 - \rho_i)$.

We pool probability loss using question weights and label-trial counts:

$$\widehat{\mathcal{L}}_{\mathrm{B}} = \frac{\sum_i v_i\{C_i(1 - q_i)^2 + (n_i - C_i)q_i^2\}}{\sum_i v_i n_i}. \tag{10}$$

Our NLL is the weighted count likelihood per label trial, $-\sum_i v_i[C_i \log q_i + (n_i - C_i)\log(1 - q_i)]/\sum_i v_i n_i$, with probabilities clipped to $[10^{-12}, 1 - 10^{-12}]$ for logarithms. Each record has

$n_i = 10$ and $v_i = 1/65$, since every question contributes thirteen model configurations and five prefixes. Each question thus has unit total record weight across its 65 model–prefix records. Our pooled metrics aggregate predictions from the separately fitted model–task readouts; subgroup summaries restrict the same weights and renormalize. We assess probability prediction with Brier and NLL, and ranking separately with AUROC.

Within-stratum ranking. For a fixed model–task–prefix-position stratum s, define positive mass $a_i = v_i C_i$ and negative mass $b_i = v_i(n_i - C_i)$. With declared ranking score z_i, we compute the count-weighted AUROC as

$$\mathrm{AUC}_s = \frac{\sum_{i,j \in s} a_i b_j \left[\mathbf{1}(z_i > z_j) + \frac{1}{2}\mathbf{1}(z_i = z_j)\right]}{\left(\sum_{i \in s} a_i\right)\left(\sum_{j \in s} b_j\right)}. \tag{11}$$

We include only strata with positive failure and success mass. We average their AUROCs using trial mass $\sum_{i \in s} v_i n_i$, with eligible-stratum coverage reported. All 195 model–task–prefix strata in the common label support have positive failure and success mass and are therefore eligible. We use the final shrunk probability as the TrajUQ ranking score. The reference baselines use the native or calibrated ranking fields specified in their fixed protocols.

Question-level bootstrap. Our full-support comparisons use 2,000 task-stratified original-question draws with seed 20260815. Our stored multiplicity matrix supplies identical draws to each method and configuration. We hold fitted predictions fixed throughout resampling.

In each draw, we assign one multiplicity to all records of an original question across models and prefix positions. Our primary nested-tuned T–M comparison uses a 95% percentile interval. The T–L comparison uses a conservative 97.5% percentile interval. For task, model, metric, and archived sensitivity comparisons, we use descriptive 95% intervals, as for the additional post-hoc L–M contrast in Appendix D.2. Our intervals quantify question-level resampling uncertainty, conditional on the selected method, fitted predictors, and analysis choices.

C.3 Independent-label outcome composition

We group the 351,000 independent label continuations by normalized-answer extraction and the evaluator's correctness judgment (Table 6). We distinguish three categories: extraction failure, extracted but incorrect, and correct. All trials contribute to the original failure-risk target and probability metrics.

Table 6: Independent-label outcome composition. Shares use the original question weights and the full 351,000-trial denominator; intervals use the retained 2,000 original-question bootstrap draws and are descriptive 95% intervals.

Outcome	Continuations	Share (%)	95% interval (%)
Parser returned no answer	21,874	6.232	[5.491,7.071]
Parsed, judged incorrect	156,712	44.65	[42.49,46.79]
Judged correct	172,414	49.12	[46.83,51.4]

To assign categories, we combine the recorded parser outcome or normalized answer with the original correctness judgment. A null normalized answer denotes an extraction failure; a recorded answer judged incorrect belongs to the second category. Format compliance is a separate attribute: 165,601 outputs satisfy the requested format, 77,399 do not, and 108,000 have no recorded format flag. A terminal-line fallback can recover a correct answer from an output that does not satisfy the format requirement.

D Additional Prediction Results

D.1 Fixed-reference results

Table 7 reports all methods and metrics; Table 8 gives Brier differences by task and model configuration. Relative to $c(O)$, our TrajUQ results give pooled $10^3\Delta$NLL of -15.58 with 95% interval

$[-16.34, -14.84]$ and ΔAUROC of $+0.06912$ with interval $[0.06590, 0.07209]$. Negative NLL and positive AUROC differences favor TrajUQ.

Table 7: Fixed-reference full metrics on the 540-question support. Values are estimates $\pm$ question-bootstrap SE; all twelve methods are retained. AUROC uses the declared baseline ranking fields and the final probabilities for the matched readouts.

Method	Brier	NLL	AUROC
Pooled			
LN-NLL	0.2495 ± 0.0004	0.6921 ± 0.0008	0.5082 ± 0.0103
EigenScore	0.2451 ± 0.0008	0.6834 ± 0.0017	0.5668 ± 0.0072
SEP-style Probe	0.2500 ± 0.0005	0.6931 ± 0.0010	0.5315 ± 0.0103
Effective Rank	0.2446 ± 0.0009	0.6824 ± 0.0017	0.5687 ± 0.0072
Sample Consistency	0.2281 ± 0.0015	0.6484 ± 0.0031	0.6086 ± 0.0087
Semantic Entropy	0.2324 ± 0.0013	0.6540 ± 0.0029	0.5847 ± 0.0102
KLE	0.2390 ± 0.0010	0.6693 ± 0.0021	0.5798 ± 0.0105
Semantic Density	0.2298 ± 0.0014	0.6523 ± 0.0030	0.5925 ± 0.0088
SRE-UQ	0.2419 ± 0.0006	0.6755 ± 0.0013	0.5388 ± 0.0120
Current-prefix Probe	0.2155 ± 0.0033	0.6244 ± 0.0077	0.5507 ± 0.0064
Matched mean $c(O)$	0.2011 ± 0.0032	0.5873 ± 0.0070	0.6160 ± 0.0076
TrajUQ	0.1947 ± 0.0030	0.5717 ± 0.0067	0.6851 ± 0.0069
MMLU-Pro			
LN-NLL	0.2498 ± 0.0006	0.6929 ± 0.0013	0.4639 ± 0.0179
EigenScore	0.2425 ± 0.0013	0.6782 ± 0.0026	0.5188 ± 0.0117
SEP-style Probe	0.2545 ± 0.0011	0.7022 ± 0.0022	0.4844 ± 0.0207
Effective Rank	0.2417 ± 0.0014	0.6766 ± 0.0029	0.5305 ± 0.0117
Sample Consistency	0.2459 ± 0.0027	0.6849 ± 0.0055	0.5364 ± 0.0152
Semantic Entropy	0.2639 ± 0.0029	0.7209 ± 0.0059	0.4722 ± 0.0218
KLE	0.2616 ± 0.0022	0.7163 ± 0.0044	0.4708 ± 0.0224
Semantic Density	0.2462 ± 0.0026	0.6855 ± 0.0054	0.5338 ± 0.0147
SRE-UQ	0.2536 ± 0.0011	0.6999 ± 0.0023	0.4560 ± 0.0235
Current-prefix Probe	0.2254 ± 0.0050	0.6475 ± 0.0117	0.5413 ± 0.0123
Matched mean $c(O)$	0.2122 ± 0.0047	0.6137 ± 0.0103	0.5871 ± 0.0153
TrajUQ	0.2055 ± 0.0045	0.5979 ± 0.0099	0.6543 ± 0.0139
StrategyQA			
LN-NLL	0.2544 ± 0.0009	0.7019 ± 0.0017	0.5065 ± 0.0209
EigenScore	0.2545 ± 0.0012	0.7023 ± 0.0025	0.5604 ± 0.0144
SEP-style Probe	0.2549 ± 0.0008	0.7029 ± 0.0016	0.5120 ± 0.0175
Effective Rank	0.2532 ± 0.0012	0.6996 ± 0.0024	0.5559 ± 0.0145
Sample Consistency	0.2335 ± 0.0033	0.6604 ± 0.0067	0.5779 ± 0.0188
Semantic Entropy	0.2337 ± 0.0020	0.6606 ± 0.0041	0.5786 ± 0.0185
KLE	0.2448 ± 0.0015	0.6845 ± 0.0030	0.5777 ± 0.0185
Semantic Density	0.2364 ± 0.0030	0.6669 ± 0.0062	0.5665 ± 0.0189
SRE-UQ	0.2451 ± 0.0009	0.6841 ± 0.0018	0.5434 ± 0.0213
Current-prefix Probe	0.2175 ± 0.0077	0.6297 ± 0.0178	0.5185 ± 0.0102
Matched mean $c(O)$	0.2073 ± 0.0075	0.6026 ± 0.0166	0.5718 ± 0.0122
TrajUQ	0.2014 ± 0.0072	0.5881 ± 0.0159	0.6540 ± 0.0113
TriviaQA			
LN-NLL	0.2442 ± 0.0006	0.6814 ± 0.0012	0.5543 ± 0.0147
EigenScore	0.2384 ± 0.0017	0.6697 ± 0.0035	0.6212 ± 0.0115
SEP-style Probe	0.2406 ± 0.0006	0.6742 ± 0.0011	0.5981 ± 0.0160
Effective Rank	0.2389 ± 0.0018	0.6709 ± 0.0037	0.6196 ± 0.0114
Sample Consistency	0.2049 ± 0.0017	0.5998 ± 0.0036	0.7116 ± 0.0112
Semantic Entropy	0.1996 ± 0.0020	0.5805 ± 0.0049	0.7032 ± 0.0106
KLE	0.2105 ± 0.0016	0.6070 ± 0.0037	0.6908 ± 0.0117
Semantic Density	0.2069 ± 0.0017	0.6044 ± 0.0035	0.6773 ± 0.0120
SRE-UQ	0.2269 ± 0.0011	0.6423 ± 0.0027	0.6169 ± 0.0169
Current-prefix Probe	0.2036 ± 0.0030	0.5960 ± 0.0078	0.5923 ± 0.0105
Matched mean $c(O)$	0.1838 ± 0.0028	0.5456 ± 0.0067	0.6891 ± 0.0109
TrajUQ	0.1771 ± 0.0028	0.5292 ± 0.0064	0.7471 ± 0.0096

Table 8: Fixed-reference TrajUQ minus $c(O)$: $10^3\Delta$Brier and ordinary 95% paired intervals. These descriptive groups share original questions across model configurations.

Group	$10^3\Delta$Brier	95% interval
Task		
MMLU-Pro	-6.785	[-7.252, -6.320]
StrategyQA	-5.918	[-6.510, -5.329]
TriviaQA	-6.680	[-7.111, -6.224]
Model configuration		
Gemma3 12B	-8.012	[-9.113, -6.964]
Gemma3 27B	-6.500	[-7.273, -5.724]
Gemma3 4B	-7.619	[-8.584, -6.694]
Llama-3.1 70B	-5.328	[-6.014, -4.642]
Llama-3.1 8B	-6.614	[-7.673, -5.628]
Qwen2.5 14B	-4.496	[-5.067, -3.967]
Qwen2.5 32B	-7.531	[-8.556, -6.491]
Qwen2.5 3B	-5.184	[-5.959, -4.439]
Qwen2.5 72B	-7.157	[-7.944, -6.311]
Qwen2.5 7B	-6.262	[-6.943, -5.623]
Yi-1.5 34B	-6.986	[-7.616, -6.330]
Yi-1.5 6B	-5.405	[-6.133, -4.693]
Yi-1.5 9B	-6.897	[-7.581, -6.209]

D.2 Nested-tuned metrics

Table 9: Nested-tuned results: estimates ± question-bootstrap SE. Task and complementary-metric comparisons are descriptive.

Readout	Brier	NLL	AUROC
Pooled			
M: optimized mean	0.2011 ± 0.003158	0.5872 ± 0.0070	0.6153 ± 0.0076
L: low-order statistics	0.2009 ± 0.003149	0.5867 ± 0.0070	0.6158 ± 0.0074
T: TrajUQ	0.1947 ± 0.003039	0.5716 ± 0.0067	0.6839 ± 0.0069
MMLU-Pro			
M: optimized mean	0.2122 ± 0.004697	0.6137 ± 0.0104	0.5870 ± 0.0153
L: low-order statistics	0.2121 ± 0.004686	0.6134 ± 0.0103	0.5865 ± 0.0148
T: TrajUQ	0.2055 ± 0.004531	0.5978 ± 0.0099	0.6550 ± 0.0139
StrategyQA			
M: optimized mean	0.2075 ± 0.007465	0.6031 ± 0.0166	0.5707 ± 0.0123
L: low-order statistics	0.2076 ± 0.007449	0.6033 ± 0.0166	0.5690 ± 0.0121
T: TrajUQ	0.2015 ± 0.007178	0.5885 ± 0.0158	0.6512 ± 0.0112
TriviaQA			
M: optimized mean	0.1836 ± 0.002871	0.5449 ± 0.0068	0.6883 ± 0.0109
L: low-order statistics	0.1830 ± 0.002860	0.5435 ± 0.0068	0.6918 ± 0.0107
T: TrajUQ	0.1770 ± 0.002779	0.5285 ± 0.0065	0.7454 ± 0.0096

Additional contrast against the optimized mean. Our post-hoc descriptive L–M contrast uses the same fixed fits and 2,000 question-bootstrap draws. We obtain a pooled L–M Brier difference of -0.000212, with 95% interval $[-0.000391, -0.000027]$, a 0.11% reduction relative to M. This contrast supplements the primary T–M and structural T–L comparisons.

D.3 Two-depth mean readout

As a complementary post-hoc analysis, we retain the average score at both observed depths. We write $m_{16} = K^{-1}\sum_{b=1}^{K} p_{b,16}$ and retain $O = K^{-1}\sum_{b=1}^{K} p_{b,64}$. Our two-depth mean reader M2 uses $K_{\mathrm{M2}} = \frac{1}{2}K_O(1 + K_{16})$, where K_{16} applies the same training-side clipping, logit transformation, standardization, and median-bandwidth rule as K_O to m_{16}. M2 uses the weighted KRR, free intercept, output clipping, and shrinkage toward O from Appendix B. Both kernels share the selected bandwidth

multiplier. Our M2 reader preserves depth-specific mean levels; T instead centers every depth across branches before representing their empirical distribution.

We fix the M2 definition and nine-candidate search before computing its evaluation results. We reuse the outer-, single-, and pair-exclusion probe-score caches and original question-grouped nested splits for fitting. Separate model–task regressors share the candidate and shrinkage selected from the pooled inner out-of-fold predictions within each outer fold, as in Appendix B.7. Our analysis covers all 35,100 prefix records from the same 540 questions, with ten branches and score depths 16 and 64.

For the post-hoc paired comparison, we match the saved nested-tuned T and M2 predictions by prefix record and score them against the same continuation outcomes with the original weights. We aggregate Bernoulli trial losses within each original question, and reuse the 2,000 task-stratified question-bootstrap draws with both predictions held fixed. Our pooled and task-specific T–M2 intervals are descriptive, unadjusted 95% percentile intervals conditional on these predictions.

Table 10: Two-depth mean reader and descriptive post-hoc comparison. The M2 columns give Brier $\pm$ SE and marginal 95% intervals. The last two columns give $\Delta = \mathrm{Brier}(T) - \mathrm{Brier}(\mathrm{M2})$ and descriptive paired 95% intervals, both multiplied by 10^3.

Task	M2 Brier $\pm$ SE	95% interval	$10^3\Delta$	Paired 95% interval
MMLU-Pro	0.2121 ± 0.004685	$[0.2034, 0.2214]$	-6.66	$[-7.142, -6.194]$
StrategyQA	0.2074 ± 0.007468	$[0.193, 0.2219]$	-5.893	$[-6.549, -5.228]$
TriviaQA	0.1836 ± 0.002851	$[0.178, 0.1891]$	-6.596	$[-7.035, -6.115]$
Pooled	0.201 ± 0.003155	$[0.1945, 0.2071]$	-6.383	$[-6.685, -6.042]$

The pooled Brier is 0.201 (Table 10). The selected bandwidth multiplier is $\beta = 2$ in every outer fold. Regularization is $\lambda = 0.001$ in fold 1 and 0.01 in the other folds; the corresponding shrinkage values in folds 0–4 are $(0.3111, 0.26, 0.3532, 0.3956, 0.3208)$.

D.4 Strict distribution-only readout

Our post-hoc control uses K_U alone, retaining the original weighted KRR, free intercept, and output clipping. Our final prediction is $(1 - \alpha_U)\mu_{\mathrm{train}} + \alpha_U q_U$, where μ_{train} is the weighted label-failure rate within the model–task training scope. Thus neither kernel nor mixture uses the prefix's O. The nine-candidate search and shared inner-OOF selection follow Appendix B.7, with training constants recomputed within every exclusion scope. We fit using outer-, single-, and pair-exclusion score caches. We fixed the definition before computing its results. For T–U, we use aligned nested-tuned predictions and the existing 2,000 task-stratified question draws, yielding descriptive, unadjusted paired 95% intervals conditional on the fits.

E Representation and Observation Ablations

E.1 Score positions

Our score-position ablations vary the coordinates supplied to U while keeping the 540-question support, $K = 10, H = 64$, endpoint mean, and independent labels fixed. We refit a nested readout and shrinkage for each position subset under the fixed-reference kernel rule, whereas the nested-tuned comparison in Section 5.4 selects among nine candidates. Removing position 64 from U retains the endpoint mean and generation horizon.

Our four-position readout has lower pooled Brier than the default two-position readout, but a three-position subset has the lowest loss (Table 11). Tables 12 and 13 quantify the paired position effects and their differences between tasks. With the endpoint mean and observation budget fixed, our comparison shows that performance depends on which depths are retained.

Table 11: Score-position results at fixed $K = 10, H = 64$ and raw endpoint mean. Each new position subset receives a nested readout refit using the original fixed kernel rule and its own shrinkage. The default and mean rows reuse their reference fits.

Retained positions	MMLU-Pro	StrategyQA	TriviaQA	Pooled
$c(O)$	0.2122	0.2073	0.1838	0.2011
8,16,32,64	0.2054	0.2013	0.1770	0.1946
16,32,64	0.2054	0.2013	0.1771	0.1946
8,32,64	0.2055	0.2013	0.1768	0.1945
8,16,64	0.2055	0.2013	0.1770	0.1946
8,16,32	0.2054	0.2013	0.1772	0.1946
Default	0.2055	0.2014	0.1771	0.1947

Table 12: Descriptive position effects: $10^3[\mathrm{Brier}(S) - \mathrm{Brier}(\{8, 16, 32, 64\})]$ with ordinary 95% paired intervals. Positive values indicate increased loss.

Retained positions	MMLU-Pro	StrategyQA	TriviaQA	Pooled
16,32,64	-0.005 [-0.118, 0.106]	+0.020 [-0.082, 0.122]	+0.168 [0.074, 0.257]	+0.061 [0.003, 0.119]
8,32,64	+0.042 [-0.056, 0.134]	-0.023 [-0.115, 0.067]	-0.197 [-0.272, -0.116]	-0.059 [-0.110, -0.007]
8,16,64	+0.055 [-0.033, 0.143]	+0.015 [-0.084, 0.107]	+0.033 [-0.055, 0.114]	+0.034 [-0.020, 0.084]
8,16,32	-0.018 [-0.165, 0.127]	-0.063 [-0.168, 0.042]	+0.285 [0.173, 0.399]	+0.068 [-0.004, 0.138]
16,64	+0.039 [-0.118, 0.187]	+0.055 [-0.105, 0.205]	+0.192 [0.075, 0.307]	+0.095 [0.014, 0.176]

Table 13: Between-task differences in position-deletion effects, multiplied by 10^3, with ordinary 95% intervals. These are descriptive, unadjusted comparisons; all deletions and task pairs are retained.

Removed	MMLU − Strategy	MMLU − Trivia	Strategy − Trivia
8	-0.025 [-0.181, 0.122]	-0.174 [-0.312, -0.030]	-0.148 [-0.288, -0.009]
16	+0.065 [-0.064, 0.189]	+0.239 [0.117, 0.356]	+0.173 [0.054, 0.295]
32	+0.040 [-0.088, 0.174]	+0.022 [-0.100, 0.148]	-0.018 [-0.144, 0.101]
64	+0.046 [-0.140, 0.223]	-0.303 [-0.488, -0.118]	-0.348 [-0.499, -0.193]

E.2 Observation width and depth

We keep the probe, readouts, and shrinkage frozen at their $K = 10, H = 64$ fits. Depth varies $H \in \{8, 16, 32, 64\}$ at $K = 10$, capped by pre-answer length; width uses the first $K \in \{1, 2, 4, 5, 8, 10\}$ branches in fixed order at $H = 64$, recomputing O and centered U.

Across H8–H64, we observe a pooled Brier decrease of 5.20% for each reader (Figure 5). The TrajUQ–mean intervals remain below zero at all four depths. Our depth sweep evaluates observation depth, truncation, and effective pre-answer length jointly, with the same H64-trained probe applied to earlier states.

We obtain lower losses for both readers as width increases. T–M intervals lie below zero at the five widths with $K \geq 2$ and include zero at $K = 1$. At $K = 1, U = \delta_0$ leaves the readers as separately fitted frozen mappings of O. At $K = 2$, the centered empirical distribution is determined by its second-moment matrix.

Table 14 reports the paired differences and uncertainty for every plotted setting.

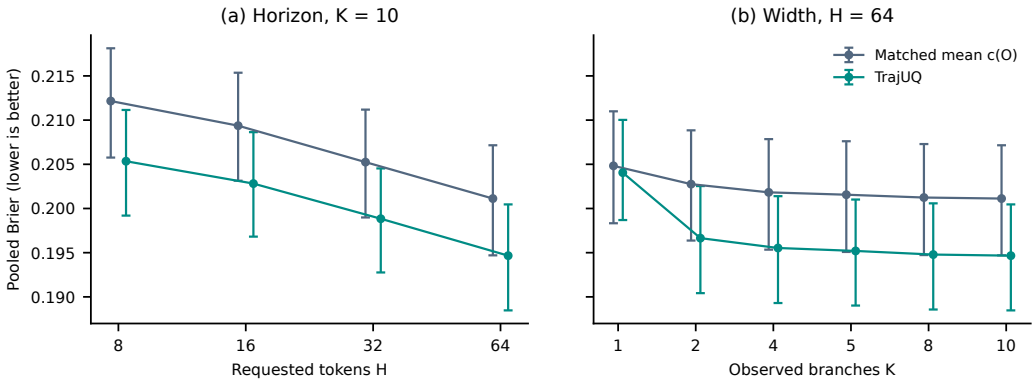

Figure 5: Frozen-reference observation sensitivity on 540 questions. Left: horizon at $K = 10$; right: width at $H = 64$. Bars give ordinary 95% question-bootstrap intervals. State-selection horizons are capped by available pre-answer branch length; width uses nested subsets. The paired differences and intervals are in Table 14.

Table 14: Frozen-reference observation sensitivity. Brier ± SE; $\Delta = TrajUQ - c(O)$ with ordinary 95% paired intervals. The effective endpoint is capped at the last available token before the final answer.

(K, H)	$c(O)$	TrajUQ	Δ [95% interval]
Horizon sensitivity: $K = 10$			
(10,8)	0.2122±0.003088	0.2053±0.002981	-0.006811 [-0.007110, -0.006514]
(10,16)	0.2094±0.003047	0.2028±0.002948	-0.006550 [-0.006828, -0.006259]
(10,32)	0.2052±0.003045	0.1988±0.002937	-0.006405 [-0.006702, -0.006111]
(10,64)	0.2011±0.003154	0.1947±0.003041	-0.006461 [-0.006759, -0.006150]
Width sensitivity: $H = 64$			
(1,64)	0.2048±0.003174	0.2041±0.003111	-0.0007722 [-0.005983, 0.005239]
(2,64)	0.2028±0.003156	0.1966±0.003064	-0.006108 [-0.006398, -0.005804]
(4,64)	0.2018±0.003150	0.1955±0.003049	-0.006296 [-0.006571, -0.006017]
(5,64)	0.2016±0.003150	0.1952±0.003044	-0.006363 [-0.006641, -0.006070]
(8,64)	0.2012±0.003155	0.1948±0.003044	-0.006452 [-0.006746, -0.006141]
(10,64)	0.2011±0.003154	0.1947±0.003041	-0.006461 [-0.006759, -0.006150]

E.3 Probe representation

Our probe study tests how altered scores perform with the reference readouts. We vary whitened PCA dimension over $\{4, 8, 16, 32, 64, 128\}$ and also use all usable stored coordinates without PCA. For each nonreference condition, we refit the upstream scaler, PCA where applicable, and logistic probe with the same question exclusions and branch-failure supervision. The downstream kernel bases, coordinate transforms, bandwidths, coefficients, and shrinkage remain frozen at the PCA32 reference. The fixed finite-coordinate mask is reused.

We obtain lower Brier with TrajUQ than the matched mean at all seven probe inputs (Table 15). With Original inputs shared by both readers, we obtain Brier 0.1908 for TrajUQ versus 0.1967 for M.

F Observation and Computational Cost

F.1 Observation accounting

Let G_b be the recorded number of generated tokens on branch b. For fixed (K, h, H), we measure observation cost as $C_{\text{obs}}(K, H) = \sum_{b=1}^{K} \min\{H, G_b\} \leq KH$, using recorded `horizon_metadata.used` in fixed branch order; pre-answer length L_b governs state selection. This observation token budget excludes prefix construction and independent full continuations for of-

Table 15: Pooled probe sensitivity. Only upstream preprocessing and the probe are refitted for nonreference inputs; downstream readouts remain frozen at PCA32. Brier ± SE and ordinary 95% paired intervals are reported.

Probe input	$c(O)$	TrajUQ	Δ [95% interval]
4	0.2063 ± 0.003084	0.1998 ± 0.002967	-0.006467 [-0.006781, -0.006142]
8	0.2038 ± 0.003124	0.1972 ± 0.003008	-0.006550 [-0.006866, -0.006234]
16	0.2020 ± 0.003167	0.1955 ± 0.003048	-0.006559 [-0.006889, -0.006231]
32	0.2011 ± 0.003154	0.1947 ± 0.003041	-0.006461 [-0.006759, -0.006150]
64	0.2003 ± 0.003205	0.1941 ± 0.003089	-0.006183 [-0.006522, -0.005835]
128	0.1992 ± 0.003255	0.1932 ± 0.003127	-0.006008 [-0.006401, -0.005595]
Original	0.1967 ± 0.003525	0.1908 ± 0.003317	-0.005979 [-0.006588, -0.005331]

fline supervision and evaluation, and differs from full collection cost and online latency. Earlier-state scoring reuses generation but may add state-access cost.

Table 16 covers all 24 configurations on the same 540 questions; Table 17 summarizes token costs and Brier by ceiling range. For width, we follow the sensitivity analysis's fixed nested branch order.

Table 16: All 24 frozen-reference configurations. Tokens denote the observation cost C_{obs} per prefix, shared by both readers; KH is an upper bound. Cells give estimates ± SE above ordinary 95% intervals.

K	H	KH	Tokens	$c(O)$ Brier	TrajUQ Brier
			7.41 ± 0.01	0.2137 ± 0.003103	0.2140 ± 0.003083
1	8	8	[7.39, 7.43]	[0.2072, 0.2198]	[0.2078, 0.2189]
			13.92 ± 0.03	0.2118 ± 0.003070	0.2113 ± 0.002964
1	16	16	[13.86, 13.98]	[0.2054, 0.2178]	[0.1992, 0.2162]
			14.82 ± 0.02	0.2128 ± 0.003093	0.2062 ± 0.002977
2	8	16	[14.77, 14.87]	[0.2063, 0.2188]	[0.2000, 0.2120]
			25.67 ± 0.08	0.2085 ± 0.003083	0.2093 ± 0.002992
1	32	32	[25.51, 25.82]	[0.2022, 0.2145]	[0.2012, 0.2131]
			27.84 ± 0.06	0.2103 ± 0.003056	0.2038 ± 0.002952
2	16	32	[27.73, 27.97]	[0.2040, 0.2163]	[0.1977, 0.2096]
			29.64 ± 0.05	0.2125 ± 0.003089	0.2058 ± 0.002981
4	8	32	[29.55, 29.73]	[0.2060, 0.2185]	[0.1996, 0.2116]
			37.05 ± 0.06	0.2123 ± 0.003085	0.2056 ± 0.002978
5	8	40	[36.94, 37.17]	[0.2059, 0.2183]	[0.1994, 0.2114]
			45.04 ± 0.22	0.2048 ± 0.003174	0.2041 ± 0.003111
1	64	64	[44.64, 45.47]	[0.1983, 0.2110]	[0.1987, 0.2100]
			51.33 ± 0.16	0.2067 ± 0.003059	0.2004 ± 0.002959
2	32	64	[51.03, 51.65]	[0.2005, 0.2127]	[0.1944, 0.2062]
			55.68 ± 0.13	0.2098 ± 0.003046	0.2033 ± 0.002948
4	16	64	[55.44, 55.92]	[0.2035, 0.2158]	[0.1973, 0.2091]
			59.28 ± 0.09	0.2121 ± 0.003086	0.2053 ± 0.002979
8	8	64	[59.10, 59.46]	[0.2057, 0.2181]	[0.1992, 0.2111]
			69.59 ± 0.16	0.2096 ± 0.003048	0.2031 ± 0.002950
5	16	80	[69.30, 69.90]	[0.2033, 0.2156]	[0.1970, 0.2088]
			74.10 ± 0.12	0.2122 ± 0.003088	0.2053 ± 0.002981
10	8	80	[73.88, 74.33]	[0.2058, 0.2181]	[0.1992, 0.2111]
			90.10 ± 0.43	0.2028 ± 0.003156	0.1966 ± 0.003064
2	64	128	[89.30, 90.98]	[0.1964, 0.2088]	[0.1904, 0.2025]
			102.6 ± 0.32	0.2059 ± 0.003050	0.1996 ± 0.002946
4	32	128	[102.1, 103.3]	[0.1996, 0.2119]	[0.1935, 0.2053]
			111.4 ± 0.25	0.2094 ± 0.003048	0.2029 ± 0.002951
8	16	128	[110.9, 111.9]	[0.2031, 0.2154]	[0.1968, 0.2086]
			128.3 ± 0.40	0.2056 ± 0.003047	0.1993 ± 0.002942
5	32	160	[127.5, 129.1]	[0.1993, 0.2116]	[0.1931, 0.2049]
			139.2 ± 0.32	0.2094 ± 0.003047	0.2028 ± 0.002948
10	16	160	[138.6, 139.8]	[0.2031, 0.2154]	[0.1968, 0.2086]

Continued on next page

Table 16 (continued)

K	H	KH	Tokens	$c(O)$ Brier	TrajUQ Brier
4	64	256	180.2±0.86 [178.6, 181.9]	0.2018±0.003150 [0.1953, 0.2078]	0.1955±0.003049 [0.1893, 0.2014]
8	32	256	205.2±0.64 [204.1, 206.5]	0.2053±0.003046 [0.1990, 0.2112]	0.1989±0.002941 [0.1928, 0.2046]
5	64	320	225.2±1.07 [223.2, 227.3]	0.2016±0.003150 [0.1951, 0.2076]	0.1952±0.003044 [0.1890, 0.2010]
10	32	320	256.6±0.80 [255.1, 258.2]	0.2052±0.003045 [0.1990, 0.2112]	0.1988±0.002937 [0.1928, 0.2045]
8	64	512	360.4±1.71 [357.2, 363.9]	0.2012±0.003155 [0.1947, 0.2073]	0.1948±0.003044 [0.1886, 0.2006]
10	64	640	450.5±2.14 [446.5, 454.9]	0.2011±0.003154 [0.1947, 0.2072]	0.1947±0.003041 [0.1885, 0.2005]

Table 17: Unweighted averages of configuration-level token and Brier estimates within each ceiling range; a descriptive summary of Table 16.

KH range	Tokens	$c(O)$ Brier	TrajUQ Brier
≤ 32	19.88	0.2116	0.2084
$(32, 64]$	49.67	0.2092	0.2037
$(64, 128]$	89.56	0.2080	0.2015
$(128, 256]$	163.2	0.2055	0.1991
$(256, 512]$	280.7	0.2027	0.1963
$(512, 640]$	450.5	0.2011	0.1947

F.2 Within-reader observation allocations

Table 18: Observation allocations with the frozen reference pipeline. Each group reports two configurations followed by their paired difference (first minus second) and 95% question-bootstrap intervals. Tokens denote the observation budget C_{obs}, shared by both readers.

(K, H) or contrast	Tokens	$c(O)$ Brier	TrajUQ Brier
A. Same token ceiling: $KH = 320$			
(5,64)	225.2	0.2016	0.1952
(10,32)	256.6	0.2052	0.1988
$(5, 64) - (10, 32)$	-31.40 [-32.12, -30.65]	-0.003686 [-0.004839, -0.002549]	-0.003644 [-0.004766, -0.002507]
B. Fewer branches at fixed depth: $H = 64$			
(8,64)	360.4	0.2012	0.1948
(10,64)	450.5	0.2011	0.1947
$(8, 64) - (10, 64)$	-90.16 [-91.04, -89.38]	+0.000116 [0.000025, 0.000207]	+0.000124 [0.000028, 0.000220]

At $KH = 320$, using five branches of up to 64 tokens, we reduce TrajUQ's Brier by 1.83% and observation tokens by 12.24% relative to ten branches of up to 32 tokens; both paired intervals favor this allocation. The mean readout also improves: equal ceilings yield different costs and losses.

At H64, with eight branches we use 20.01% fewer tokens than ten, with 0.06% higher T Brier and a small positive paired interval. These frozen-pipeline tradeoffs complement Table 17's descriptive grid averages.

F.3 Offline time and memory measurements

We measure runtime for Qwen2.5-7B on all three tasks in outer fold 0, with 36 evaluation questions per task, on an Intel Core i9-13900HX running Windows 11 with two BLAS threads. Probes use

the frozen 32-coordinate pipeline; mapping arms share score pairs and retain fitted coefficients and training bases. Batch sizes, stages, and repeats are specified below.

Table 19 distinguishes CPU-stage and worker wall times. In both measurements, we exclude generation and process startup; wall time starts after NumPy import and specification reading. We load inputs separately in probe workers, and basis loading reconstructs fitted states. We use the operating system's existing file-cache state. We sample RSS every 50 ms and at worker start/end. Distinct worker inputs make wall times and memory peaks nonadditive.

Table 19: Offline CPU cost for the frozen Qwen2.5-7B readouts, with 180 evaluation prefixes and ten branches per task. Values are means ± sample standard deviations over three fresh-worker repeats. The operation column gives the named operation time divided by 180; worker wall includes setup and loading within the worker but excludes process startup. RSS is the sampled absolute resident peak; increment is relative to that worker's starting RSS. The three tasks and seven stage/arm combinations give 63 worker measurements.

Stage	Operation (ms/prefix)	Worker wall (s)	Peak RSS (MiB)	Increment (MiB)
MMLU-Pro				
Read archived grids	42.23±32.33	7.823±5.863	94.33±0.65	56.64±0.60
Endpoint probe ($H = 64$)	0.953±0.014	8.526±1.735	203.1±0.57	165.4±0.49
Earlier probe ($h = 16$)	0.952±0.068	8.468±2.244	202.5±1.28	164.7±1.24
$c(O)$: load frozen basis	0.186±0.023	0.488±0.006	59.89±0.17	22.20±0.21
TrajUQ: load frozen basis	0.164±0.007	0.481±0.010	60.11±0.32	22.49±0.36
$c(O)$: probability readout	0.125±0.007	0.503±0.058	60.43±0.46	22.74±0.49
TrajUQ: probability readout	4.990±0.150	1.352±0.012	62.60±0.44	24.90±0.35
StrategyQA				
Read archived grids	43.36±31.12	8.002±5.576	94.53±0.16	56.83±0.11
Endpoint probe ($H = 64$)	0.859±0.030	7.808±1.808	203.7±1.76	166.0±1.75
Earlier probe ($h = 16$)	0.889±0.026	8.138±1.401	202.9±1.23	165.2±1.22
$c(O)$: load frozen basis	0.169±0.038	0.462±0.034	59.88±0.06	22.20±0.10
TrajUQ: load frozen basis	0.154±0.007	0.442±0.021	59.99±0.22	22.34±0.34
$c(O)$: probability readout	0.112±0.008	0.516±0.012	60.40±0.32	22.74±0.22
TrajUQ: probability readout	4.754±0.514	1.338±0.084	63.11±0.29	25.46±0.28
TriviaQA				
Read archived grids	43.94±34.65	8.093±6.248	94.39±0.98	56.79±0.99
Endpoint probe ($H = 64$)	0.881±0.043	7.967±1.955	202.6±1.52	164.9±1.52
Earlier probe ($h = 16$)	0.860±0.020	7.956±2.114	202.9±0.81	165.1±0.80
$c(O)$: load frozen basis	0.192±0.029	0.516±0.034	59.96±0.24	22.33±0.27
TrajUQ: load frozen basis	0.167±0.010	0.492±0.012	60.01±0.06	22.33±0.11
$c(O)$: probability readout	0.116±0.015	0.533±0.051	60.16±0.06	22.44±0.02
TrajUQ: probability readout	4.983±0.065	1.387±0.020	62.95±0.16	25.25±0.22

At fixed (K, H), we share generation across both readers; T additionally scores earlier states and computes distribution similarities. Across tasks, our offline CPU probability mapping for Qwen2.5-7B takes 4.754–4.990 ms per prefix for T versus 0.112–0.125 ms for $c(O)$ (Table 19), separately from generation and end-to-end latency.

G Post-hoc Outcome and Screening Analyses

G.1 Outcome losses and offline screening

Figure 6 summarizes two post-hoc nested-tuned diagnostics. Most of our T–M Brier reduction occurs among parsed incorrect and correct answers.

At prespecified 50% coverage within model–task–prefix strata, we select lower empirical failure-rate prefixes with T than M and M2 (Table 21).

For M, T, and M2, we use the same 35,100 prefixes from 540 questions, 351,000 continuation outcomes, original question weights, and outer folds. We hold nested-tuned predictions and judgments

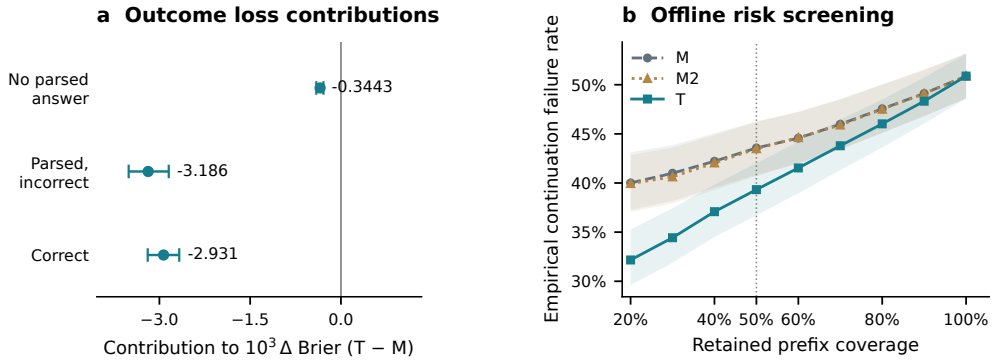

Figure 6: Post-hoc outcome and screening diagnostics under the nested-tuned protocol on 540 questions. Left: contributions to the overall T–M Brier difference, using a common full-label denominator; negative values favor T. Right: empirical continuation failure rates after retaining low-risk prefixes at each coverage. Bars and bands use the same 2,000 question-level draws and show descriptive pointwise 95% intervals. The dashed line marks the prespecified 50% comparison.

from the readout comparison fixed across 2,000 task-stratified question draws (Appendix C.2), with one multiplicity across each question's models, prefixes, and continuations.

Outcome loss composition. The three classes in Appendix C.3 partition all label trials. We compute conditional Brier by dividing each class's weighted loss by its trial weight; for its T–M contribution, we divide the class's weighted loss difference by the common full-label weight. Contributions sum to the overall difference, descriptively decomposing loss by evaluator outcome.

Table 20: Outcome composition of the T–M loss difference. Shares use the original weights; contribution estimates and their intervals are multiplied by 10^3. Intervals are descriptive paired 95% intervals.

Outcome	Share (%)	M Brier	T Brier	Contribution	95% interval
No parsed answer	6.232	0.1664	0.1608	-0.3443	$[-0.4068, -0.2891]$
Parsed, incorrect	44.65	0.2029	0.1957	-3.186	$[-3.507, -2.845]$
Correct	49.12	0.2039	0.198	-2.931	$[-3.193, -2.672]$

The classes contain 21,874, 156,712, and 172,414 trials, covering all 351,000 outcomes.

Equal-coverage screening. Within model–task–prefix strata, we retain a fraction of original trial weight in ascending order of predicted risk; boundary ties receive common, label-independent fractional inclusion. We evaluate coverage at 20%, 30%, . . . , 100%, with T–M at 50% designated primary before computation. We reapply selection in each question draw with resampled weights. Pooled rates divide selected independent-label failure mass by selected trial weight; at 100%, all readers return 50.88%.

Table 21: Post-hoc offline risk screening at 50% coverage. Rates and marginal intervals are percentages; contrasts and their paired intervals are percentage points. T–M2 is descriptive.

Reader or contrast	Estimate	95% interval
M	43.56	$[40.81, 46.26]$
M2	43.48	$[40.71, 46.22]$
T	39.33	$[36.71, 42.03]$
T–M	-4.239	$[-4.581, -3.798]$
T–M2	-4.159	$[-4.471, -3.692]$

Our analysis (Figure 6) covers offline selection under an unchanged generation policy. Rates use finite sampled continuations; curve bands and class intervals are descriptive, pointwise, and unadjusted.

H Cross-budget Prediction and Screening

H.1 Post-hoc comparison family and uncertainty

Our fixed-reference grid covers the default comparison's 35,100 evaluation records, independent outcomes, and question weights. For each allocation $j = (K, H)$, $\Delta_j = \mathrm{Brier}(T_j) - \mathrm{Brier}(M_{10,64})$. Table 22 reports point contrasts for the 20 multi-branch allocations with $K \geq 2$. Our simultaneous intervals retain the original 24-allocation calibration.

For the original calibration, we hold the fitted predictions fixed across 2,000 task-stratified question-bootstrap draws (Appendix C.2). For paired difference $\Delta_j^{*(b)}$ and its across-draw standard deviation s_j, the empirical 95th percentile of $\max_j |\Delta_j^{*(b)} - \Delta_j|/s_j$ over that 24-allocation family gives $c_{.95} = 2.706$. We retain the intervals $\Delta_j \pm c_{.95} s_j$. With the same weights and paired draws, two branches consume 20.00% of reference tokens (descriptive 95% ratio interval [19.97%,20.02%]).

Table 22: Post-hoc T(K, H) versus M(10,64) for the 20 multi-branch allocations with $K \geq 2$. The reference has Brier 0.2011 and 450.5 observation tokens. Differences and simultaneous intervals are multiplied by 10^3; negative differences favour T. Intervals retain the original 24-contrast calibration.

K	H	Tokens	T Brier	$10^3\Delta$	Simultaneous 95% interval
2	8	14.82	0.2062	5.03	[1.883, 8.176]
2	16	27.84	0.2038	2.654	[0.115, 5.192]
2	32	51.33	0.2004	-0.7058	[-2.359, 0.9473]
2	64	90.1	0.1966	-4.478	[-5.133, -3.823]
4	8	29.64	0.2058	4.643	[1.416, 7.87]
4	16	55.68	0.2033	2.221	[-0.3789, 4.821]
4	32	102.6	0.1996	-1.515	[-3.221, 0.1902]
4	64	180.2	0.1955	-5.593	[-6.087, -5.1]
5	8	37.05	0.2056	4.457	[1.213, 7.702]
5	16	69.59	0.2031	1.951	[-0.6669, 4.57]
5	32	128.3	0.1993	-1.871	[-3.595, -0.1469]
5	64	225.2	0.1952	-5.928	[-6.396, -5.46]
8	8	59.28	0.2053	4.209	[0.9427, 7.475]
8	16	111.4	0.2029	1.754	[-0.8705, 4.378]
8	32	205.2	0.1989	-2.226	[-3.958, -0.4943]
8	64	360.4	0.1948	-6.337	[-6.769, -5.904]
10	8	74.1	0.2053	4.223	[0.9426, 7.503]
10	16	139.2	0.2028	1.695	[-0.944, 4.333]
10	32	256.6	0.1988	-2.284	[-4.012, -0.5554]
10	64	450.5	0.1947	-6.461	[-6.882, -6.04]

H.2 Task and model-configuration summaries

Table 23 gives descriptive ordinary 95% paired intervals for T(2,64) minus M(10,64). Our model configurations share the same questions, and we retain all records of each question together in each bootstrap draw. Qwen2.5-14B's point difference favours T, and its interval includes zero.

H.3 Low-budget offline screening

Our post-hoc comparison applies Appendix G's weighted, stratified screening and tie rules to fixed-reference T(2,64) and M(10,64), separately from the nested-tuned default-budget analysis, with

Table 23: Two-branch T versus ten-branch M by task and model configuration. Brier differences and descriptive 95% intervals are multiplied by 10^3.

Group	$10^3\Delta$	Paired interval
MMLU-Pro	-5.446	[-6.191, -4.687]
StrategyQA	-5.149	[-5.813, -4.501]
TriviaQA	-2.839	[-3.759, -1.854]
Gemma3-12b	-6.739	[-8.063, -5.351]
Gemma3-27b	-5.465	[-7.029, -3.947]
Gemma3-4b	-4.862	[-6.518, -3.233]
Llama-3.1-70b	-1.994	[-3.804, -0.1613]
Llama-3.1-8b	-4.741	[-6.16, -3.274]
Qwen2.5-14b	-1.507	[-3.124, 0.1483]
Qwen2.5-32b	-5.477	[-6.886, -4.039]
Qwen2.5-3b	-4.064	[-5.334, -2.817]
Qwen2.5-72b	-4.456	[-6.015, -2.808]
Qwen2.5-7b	-5.188	[-6.377, -3.986]
Yi-1.5-34b	-4.096	[-5.881, -2.279]
Yi-1.5-6b	-4.103	[-5.005, -3.189]
Yi-1.5-9b	-5.519	[-6.692, -4.252]

generation unchanged. In each draw, we reselect using resampled weights; failure rates divide selected independent-label failure mass by selected trial mass.

We evaluate coverage at 20%,30%,...,100%, with 50% designated as the focus before computation. Intervals are descriptive and pointwise rather than simultaneous.

Table 24: Fixed-reference low-budget screening. Failure rates are percentages; differences and descriptive 95% paired intervals are percentage points. The methods coincide at full coverage.

Coverage	M(10,64)	T(2,64)	T–M	Paired interval
20%	39.8	33.15	-6.651	[-7.521, -5.526]
30%	40.92	35.9	-5.02	[-5.655, -4.096]
40%	42.13	38.16	-3.976	[-4.614, -3.452]
50%	43.48	39.97	-3.517	[-3.918, -2.958]
60%	44.59	42.16	-2.423	[-2.774, -1.957]
70%	46.01	44.4	-1.608	[-1.842, -1.278]
80%	47.51	46.44	-1.073	[-1.325, -0.8878]
90%	49.12	48.45	-0.6654	[-0.7949, -0.515]
100%	50.88	50.88	0	[0, 0]

I OFFLINE CANDIDATE SELECTION ON TRIVIAQA

TriviaQA provides an open-domain factual question-answering case study with varied answer expressions. We use it for a post-hoc analysis of offline selection from saved candidates. Our analysis includes all 180 TriviaQA questions in the evaluation set. For each model, our analysis uses ten archived feature candidates from the 25% anchor prefix, giving 2,340 sets across thirteen model configurations. We return one answer from every set, including identical-answer sets, using the original correctness judgments.

At each observation checkpoint, all ten candidates remain available: we update their scores and choose the candidate with the lowest score. The final checkpoint determines which saved answer is returned. Let $F(O, U)$ be the complete frozen reference readout, including shrinkage. The original question determines the outer test fold from which we load the frozen probe, readout, and shrinkage; fitting and parameter selection exclude that question. At checkpoint h, we assign the candidate score $s_{b,h} = F(p_{b,h}, U_{\mathrm{pool,h}})$, where the common distribution centers the ten score vectors at positions $\min(16, h)$ and h; ties follow numeric branch ID. We substitute the candidate score for the group mean to obtain a ranking score for answer selection. The checkpoints are 8, 16, 32, and 64 tokens, with the reported outcome using the final choice at 64 tokens. We define the single-answer baseline as a uniform draw from the same ten candidates; its expected failure is their mean failure label.

GUARD selects short branches by mean token entropy (Zhu et al., 2026). Our GUARD-style selection adapts this rule to the saved candidate pool, selecting the continuation with the lowest average token entropy. We compute the saved average from generation-processed token distributions over the full continuation, including any later text and final answer; TrajUQ uses scores at the observation checkpoints. The failure-rate difference, TrajUQ minus GUARD-style selection, is -2.436 percentage points, with descriptive paired 95% interval $[-3.761, -1.154]$.

Our separately fitted Current-prefix Probe from Table 2 scores each candidate state at the same checkpoints, with the same branch-ID tie rule and final-checkpoint selection. Its failure rate is 55.47%, with a difference of -1 percentage point from Single answer and a paired 95% interval of $[-2.166, -0.003]$.

Our Branch probe baseline scores each candidate as $s^{\mathrm{probe}}_{b,h} = p_{b,h}$ and selects the minimum, using the same checkpoints and branch-ID tie rule as TrajUQ (T). With Branch probe, we obtain 53.16% failure, a difference of -3.31 percentage points from Single answer (paired 95% interval: $[-4.358, -2.401]$). TrajUQ lowers failure by approximately 1.24 percentage points relative to Branch probe; their direct paired 95% interval is $[-1.853, -0.819]$.

We average rates equally across questions and model configurations. We compute the intervals for GUARD-style selection and T versus Single answer, and for T versus GUARD-style selection, using 2,000 paired bootstrap draws over the 180 original questions, shared across all thirteen model configurations. These bootstrap intervals are descriptive and unadjusted for selection across tasks or observation settings. TrajUQ has lower failure point estimates than single-answer sampling in ten of thirteen model configurations and than GUARD-style selection in eight (supplement: `results/triviaqa_selection_by_model.csv`).

Table 25: Final-answer failure on TriviaQA after selection from ten fixed candidates (180 questions, thirteen models). Failure rates are percentages; differences from the single-answer baseline and descriptive paired 95% intervals are percentage points.

Selection	Failure	Difference	Paired interval
Single answer	56.47	–	–
Current-prefix Probe	55.47	-1.000	[-2.166, -0.003]
GUARD-style selection	54.36	-2.111	[-3.073, -1.137]
Branch probe	53.16	-3.310	[-4.358, -2.401]
TrajUQ	51.92	-4.547	[-5.547, -3.568]