# RELIABILITY-AWARE BAYESIAN SELF-CONSISTENCY FOR EVIDENCE AGGREGATION AND RISK-CONTROLLED ABSTENTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Self-consistency (SC) selects the modal answer among sampled language-model traces, but vote support alone does not express whether heterogeneous evidence sources are reliable for the current question. We introduce *Reliability-Aware Bayesian Self-Consistency* (RA-BSC), a categorical evidence model that jointly infers the correct candidate and question-dependent trust in generators, answer verifiers, and process verifiers. A Laplace approximation propagates uncertainty over answer probabilities and source-trust functions, while a disjoint Learn-Then-Test (LTT) split supplies the selective-risk guarantee. We evaluate 77,940 traces from three Qwen2.5 generators on ARC-Challenge, CommonsenseQA, and StrategyQA using a training-split estimated candidate-miss prior. Controlled ablations show that structured reliability improves evidence aggregation, whereas Bayesian marginalization contributes uncertainty-aware selective information rather than universal point-prediction gains. RA-BSC improves over Bayesian softmax on high-conflict ARC and CommonsenseQA questions, but fixed or adaptive MAP reliability can be stronger for accuracy; verifier-weighted voting and classical calibration also remain competitive. A focused ARC audit that replaces one Qwen generator with Gemma-2B preserves this distinction: fixed reliability has the highest accuracy, while the RA-BSC lower-quantile score has the best AURC and LTT coverage. RA-BSC attains nonempty LTT certificates on ARC and CommonsenseQA, while StrategyQA remains uncertifiable at the target risk. These results separate three objectives that are often conflated: answer aggregation, scalar calibration, and certifiable abstention.

## 1 INTRODUCTION

Self-consistency samples reasoning traces and returns their modal answer (Wang et al., 2023). Its winning vote fraction is often used as confidence, despite being only a summary of repeated samples. It does not express disagreement between generators, external evidence about a candidate answer, or whether any one evidence source is trustworthy on the current question.

We study trace post-processing when sampled answers are available but generator logits need not be released. Auxiliary verifiers provide additional evidence for candidate answers, while confidence-informed SC uses internal generator confidence to weight paths (Taubenfeld et al., 2025). These signals can conflict: a generator may agree with the vote but a verifier may reject the associated reasoning, or an answer verifier may disagree with a process verifier. A fixed regression coefficient treats each channel as having the same reliability on every question. Our question is whether the reliability of the evidence itself should be inferred.

We introduce RA-BSC, a reliability-aware Bayesian categorical evidence aggregator. All answers extracted for one question compete in a multinomial model with an explicit null state. In addition to candidate evidence, the model has latent source-trust functions whose logits depend on low-dimensional conflict features for that question. It can therefore express posterior statements such as "the answer verifier is less trusted than the process verifier on this item" instead of assigning one fixed global weight to every source. A Gaussian prior, MAP fit, and Laplace approximation yield posterior draws over candidate probabilities and source trust. The posterior mean selects an answer



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

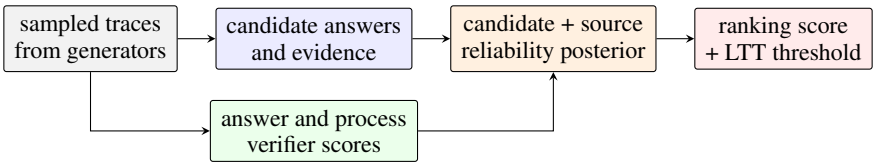


Figure 1: RA-BSC jointly infers candidate probabilities and question-dependent source reliability. LTT, not the Bayesian posterior alone, supplies the selective-risk statement.

and supplies predictive confidence; posterior lower quantiles are used only as selective ranking scores, not as calibrated probabilities. A separate LTT layer makes the risk-control claim. Crucially, candidate features are used only to define the observed evidence and question context; they are not counted again as a second likelihood term.

**Contributions.** (i) We formulate SC post-processing as categorical evidence aggregation with an explicit, training-estimated candidate-miss prior. (ii) We introduce question-dependent trust functions for generators and verifiers and propagate parameter uncertainty into answer and source-trust posteriors. (iii) A controlled staircase—fixed reliability, adaptive MAP, and posterior-marginalized RA-BSC—together with Hessian and source-duplication audits separates the effect of structured reliability from Bayesian integration. (iv) Across 77,940 traces and 20 question-level splits, we show that aggregation, scalar calibration, and certifiable abstention are distinct objectives, identify both the conflict regimes where reliability modeling helps and those where simpler baselines remain stronger, and test the resulting pattern in a focused cross-family ARC audit.

## 2 Related Work

**Sampled reasoning and verifiers.** Chain-of-thought prompting and SC established inference-time sampling for reasoning (Wei et al., 2022; Wang et al., 2023). CISC uses generator-internal confidence to weight sampled paths (Taubenfeld et al., 2025). Generative verifiers improve Best-of-$N$ selection by scoring candidate solutions (Zhang et al., 2025). Our external verifiers are observable channels in a candidate-level fusion model; unlike fixed verifier-weighted voting, RA-BSC represents their trustworthiness as uncertain and question-dependent.

**Calibration and risk control.** Platt and beta calibration transform scalar confidence after answer selection (Platt, 1999; Kull et al., 2017); they are strong baselines here. Selective prediction and LTT calibrate a held-out decision threshold to control risk (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2022; Bates et al., 2021). Sample-then-filter frameworks connect sampling with controlled filtering (Wang et al., 2025; 2026). Bayesian model averaging combines predictive distributions under posterior model weights (Hoeting et al., 1999); related Bayesian SC and stopping formulations primarily represent uncertainty over candidate correctness or stopping decisions. Our distinction is to model uncertainty over the reliability of heterogeneous evidence sources themselves, while retaining LTT as the distribution-free decision layer.

## 3 Reliability-Aware Bayesian Evidence Aggregation

### 3.1 Candidate set and categorical competition

For question $x$, let $\mathcal{A}(x)$ be the distinct normalized answers extracted from traces and let $K_x = |\mathcal{A}(x)|$. The observed evidence consists of per-generator vote fractions, answer-verifier scores, process-verifier scores, and low-dimensional question-level conflict features. These quantities are modeled as source observations below; they are not reused as a feature-based candidate prior. Gold labels are excluded from candidate construction and verifier prompts.

Exactly one semantic answer is correct, but it may be absent from $\mathcal{A}(x)$. We augment the candidate set by $\varnothing$. Rather than assigning the null state mass mechanically because it is an additional category, we estimate its prior from training-split candidate recall. For training question $i$, let $M_i = \mathbb{I}\{y_i^{\text{gold}} \notin$

$\mathcal{A}(x_i)\}$. With a weak $\mathrm{Beta}(1,1)$ smoothing prior, define

$$\widehat{\pi}_{\varnothing} = \frac{1 + \sum_{i \in \mathcal{D}_{\mathrm{train}}} M_i}{2 + |\mathcal{D}_{\mathrm{train}}|}, \qquad p_0(y = \varnothing \mid x, \mathcal{D}_{\mathrm{train}}) = \widehat{\pi}_{\varnothing}, \quad p_0(y = a \mid x, \mathcal{D}_{\mathrm{train}}) = \frac{1 - \widehat{\pi}_{\varnothing}}{K_x}. \tag{1}$$

The prior is computed separately within each training split and uses no calibration or test labels. It is independent of the evidence vector $E$ used in the source likelihood. The null state therefore represents the estimated probability that the correct answer was missed by candidate extraction, rather than a mechanical $1/(K_x + 1)$ allocation. It absorbs posterior mass when the correct answer is absent, is never emitted as a semantic answer, and its mass is retained in predictive confidence and abstention. Final prediction is the posterior-mean argmax over $\mathcal{A}(x)$; abstention is determined by a posterior score and LTT.

### 3.2 Latent source reliability

Let $j \in \mathcal{J}$ index the three generators, answer verifier, and process verifier. For candidate $a$, let $e_{ja} \in [0, 1]$ denote source support or score. The implementation uses the three-dimensional question summary

$$h(E) = \left( \frac{1}{K_x} \sum_{a \in \mathcal{A}(x)} d_a, \ \max_{a \in \mathcal{A}(x)} v_a^{\mathrm{mean}}, \ \left| \max_a q_a^{\mathrm{ans}} - \max_a q_a^{\mathrm{proc}} \right| \right), \tag{2}$$

where, for generator-specific vote fractions $v_{ga}$ and $\bar{v}_a = G^{-1} \sum_g v_{ga}$, the stored disagreement value is

$$d_a := -\frac{\sum_{b \in \mathcal{A}(x)} \bar{v}_b \log(\max\{\bar{v}_b, 10^{-12}\})}{\log K_x} \quad (K_x > 1), \tag{3}$$

with $d_a := 0$ when $K_x = 1$. This question-level normalized entropy is repeated on each candidate row, exactly as in the released candidate constructor; $v_a^{\mathrm{mean}} = \bar{v}_a$, and $q_a^{\mathrm{ans}}$ and $q_a^{\mathrm{proc}}$ are the two verifier scores. Each coordinate is standardized using the training-split mean and standard deviation (a zero-variance scale is set to one). We introduce source parameters $\theta_j = (\mu_j, \gamma_j)$ and define the conditional reliability function

$$r_j(E; \theta_j) = \sigma(\mu_j + \gamma_j^\top h(E)), \qquad \theta_j \sim \mathcal{N}(0, \lambda^{-1} I). \tag{4}$$

Conditional on a candidate answer $y = a$, source $j$ emits a categorical evidence token $c \in \mathcal{A}(x)$ from a correct-support distribution with probability $r_j(E; \theta_j)$ and from an uninformative categorical background with probability $1 - r_j(E; \theta_j)$:

$$p_j(c \mid y = a, E, r_j) = r_j(E; \theta_j) \mathbb{I}\{c = a\} + \frac{1 - r_j(E; \theta_j)}{K_x}, \qquad a \in \mathcal{A}(x). \tag{5}$$

For the explicit null state, the observed candidate evidence is uninformative regardless of the reliability parameter:

$$p_j(c \mid y = \varnothing, x) = \frac{1}{K_x}, \qquad c \in \mathcal{A}(x). \tag{6}$$

Thus $r_j = 0$ means that a source provides no information about which extracted candidate is correct; it does not mean that the source is certainly wrong. The null distribution is normalized even though the correct answer is outside the extracted candidate set. The observed source vector $e_j$ is a normalized fractional count over candidates, so we use the following fractional categorical log likelihood:

$$\ell_j(y \mid e, E, \theta_j) = \sum_{c \in \mathcal{A}(x)} e_{jc} \log \left[ p_j(c \mid y, E, r_j(E; \theta_j)) + \epsilon \right]. \tag{7}$$

The source-parameter posterior is learned from training questions. There $y_i$ is observed from the benchmark label: it is the gold candidate when that label appears in $\mathcal{A}(x_i)$ and is $\varnothing$ otherwise. Writing $E_i$ for all normalized source vectors of question $i$, the training posterior is

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{train}}) \propto p(\boldsymbol{\theta}) \prod_{i \in \mathcal{D}_{\mathrm{train}}} \prod_{j \in \mathcal{J}} p_j(e_{ij} \mid y_i, E_i, \theta_j), \tag{8}$$

where $p_j(e_j \mid y, E, \theta_j) = \exp\{\ell_j(y \mid e_j, E, \theta_j)\}$ under the fractional-count likelihood and $\ell_j(\varnothing \mid e, E) = \sum_c e_{jc} \log(1/K_x)$ follows Eq. 6. The fitted parameters define $q(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})$, approximated by a diagonal observed-Hessian Laplace posterior around the training MAP:

$$q(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}}) = \mathcal{N}\left(\widehat{\boldsymbol{\theta}}_{\text{MAP}}, \operatorname{diag}(H)^{-1}\right), \qquad H = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\theta}}_{\text{MAP}}), \tag{9}$$

The reliability posterior is the push-forward of this parameter posterior through $r_j(E; \theta_j)$. This factorization makes the probabilistic story explicit: candidate prior $\times$ reliability-weighted source likelihood $\rightarrow$ posterior answer. The same vote or verifier observation is therefore not counted in both a feature-based candidate softmax and a source likelihood. We optimize the negative log posterior on training questions. For a test item, $y$ is latent and no global parameter is updated; prediction uses

$$p(y \mid E, x, \mathcal{D}_{\text{train}}) \propto p_0(y \mid x, \mathcal{D}_{\text{train}}) \int \left[\prod_{j \in \mathcal{J}} p_j(e_j \mid y, E, \theta_j)\right] q(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})\, d\boldsymbol{\theta}. \tag{10}$$

No test item updates the global source parameters. Strictly, $r_j(E)$ is not an independently sampled per-question Beta variable; it is a question-dependent latent trust weight induced by uncertainty over $(\mu_j, \gamma_j)$. Thus this is a conditional/generalized evidence model given observed conflict summaries $h(E)$, rather than a claim that those summaries are generated independently of the evidence. This compact MAP-plus-Laplace construction is the additional Bayesian object, not a claim that source accuracy is observed without uncertainty. LTT remains the only finite-sample risk-control layer.

We predict $\hat{y} = \arg\max_{a \in \mathcal{A}(x)} \mathbb{E}_q p(y = a \mid E, x)$ over extracted candidates. The RA-BSC reliability-aware score used for selective ranking is the unconditioned posterior-mean probability assigned to this winning candidate,

$$s_{\text{RA}}(x) = \mathbb{E}_q\left[p(y = \hat{y} \mid E, x, \boldsymbol{\theta})\right]. \tag{11}$$

Here $s_{\text{RA}}$ retains posterior mass assigned to the null state, so missing-candidate uncertainty lowers the score and can trigger abstention; it is not a finite-sample guarantee. We additionally evaluate an RA-specific conservative score using the same joint Laplace draws,

$$s_{\text{RA},.10}(x) = q_{.10}(p(y = \hat{y} \mid E, x, \boldsymbol{\theta}))\,. \tag{12}$$

The posterior mean remains the predictive confidence used for calibration; Eq. 12 is used only for selective ranking and LTT. The null state is never emitted as a semantic answer. Abstention is determined by the selected score and the separately calibrated LTT threshold.

### 3.3 Risk-controlled abstention

Questions are split into disjoint 20% training, 30% calibration, and 50% test partitions. For each fixed score, we test $M = 21$ thresholds $\mathcal{T} = \{0, .05, \ldots, 1\}$. With $E_j$ calibration errors among $N_j$ answered questions,

$$\overline{R}_j = \text{Beta}^{-1}\left(1 - \frac{\delta}{M}; E_j + 1, N_j - E_j\right), \quad \delta = 0.05. \tag{13}$$

We retain thresholds with $\overline{R}_j \leq \alpha = 0.20$ and choose the retained threshold with maximum calibration coverage. When $N_j = E_j$, the upper bound is defined as $\overline{R}_j = 1$, the limiting Clopper–Pearson value. Bonferroni gives simultaneous $0.95$ coverage for the fixed threshold family. If no threshold is retained, the system abstains on every test question.

## 4 Experimental Setup

We collect 20 temperature-$0.7$ samples per question from Qwen2.5-1.5B, 3B, and 7B Instruct (Qwen Team, 2024). The collection contains 77,940 traces: the full prepared set of 299 ARC-Challenge questions (Clark et al., 2018) and deterministic first-500 subsets of CommonsenseQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021), fixed before model fitting to bound generation and verification cost. Each question has 60 traces. All methods share parsing, answer normalization, splits, and tie breaking.

As a focused generator-family audit, we repeat ARC-Challenge after replacing Qwen2.5-1.5B with Gemma-2-2B-Instruct (Gemma 2 Team, 2024), retaining Qwen2.5-3B and 7B. This gives 17,940 traces over the same 299 questions, with 20 samples from each generator. We rebuild candidate sets and both verifier channels for this mixture and rerun the same 20 splits without changing hyperparameters or selecting a method from its test outcomes.

For RA-BSC, each generator contributes its per-candidate vote fraction (the number of its 20 traces supporting the candidate divided by 20). Answer- and process-verifier scores are converted separately into candidate distributions as $e_{ja} = q_{ja} / \sum_b q_{jb}$. All five source vectors therefore sum to one over $\mathcal{A}(x)$; these are the fractional counts in Eq. 7. A verifier score is not silently interpreted as a probability of correctness. The null prior in Eq. 1 is estimated only from training candidate recall and does not reuse these observations.

Both verifier channels use Microsoft Phi-3-mini-4k-instruct in bfloat16 with teacher-forced label scoring. The answer verifier receives the question, options, and proposed answer; the process verifier additionally receives sampled supporting rationales, truncated to 48 words per generator. Both prompts request exactly "yes" or "no", and $q_a = \exp L_{\text{yes}} / (\exp L_{\text{yes}} + \exp L_{\text{no}})$ is computed from continuation log probabilities. Prompts are capped at 256 tokens; there is no sampling temperature because scoring is teacher-forced. Prompt constructors and model identifiers are included in the artifact.

Table 4 reports the candidate-recall audit. For each split, the training miss rate determines $\widehat{\pi}_{\varnothing}$ through Eq. 1; calibration and test recalls are reported descriptively and are never used to fit the posterior or select a score.

Besides pooled majority, we evaluate verifier-weighted SC, generator-confidence weighted SC, deterministic MAP softmax, Bayesian posterior-mean softmax, deterministic evidence fusion with the same observed inputs, fixed-reliability BSC, adaptive-MAP BSC, and RA-BSC. Verifier-weighted SC assigns candidate mass proportional to $v_a^{\text{mean}} \sqrt{q_a^{\text{ans}} q_a^{\text{proc}}}$ and normalizes over candidates. Deterministic evidence fusion is a regularized multinomial logistic comparator whose candidate design concatenates the five source values with all source-by-$h(E)$ interactions; it uses the same training-estimated null prior but no posterior integration. The generator-confidence baseline assigns $w_r = \exp(\bar{\ell}_r / T)$ with $T = 1$; weights are normalized within each question and generator before generator-level answer masses are averaged. We call this a *within-generator-normalized CISC-style oracle-information diagnostic*, not an exact CISC reproduction. Platt and beta calibration are fitted only on the 20% training split; the disjoint 30% calibration split is reserved for choosing the LTT threshold. Optimization, prior, numerical-stability, Hessian, and Monte Carlo settings are specified in Appendix A.1 and the released artifact.

For every predictive method, $c_i$ is confidence assigned to final prediction $\hat{y}_i$ and $r_i = \mathbb{I}\{\hat{y}_i = y_i\}$. With 15 equal-width bins, $\text{ECE} = \sum_b (n_b/n) |\bar{c}_b - \bar{r}_b|$ and $\text{Brier} = n^{-1} \sum_i (c_i - r_i)^2$. Thus accuracy, ECE, and Brier use predictive confidence, while posterior lower quantiles are only selective scores. Classical calibration results are reported in Table 8. We also report AURC, LTT coverage/risk, source-reliability summaries, and nonempty-certificate and target-risk-exceedance frequencies over 20 splits.

# 5 Results

## 5.1 Structured reliability improves evidence aggregation

Table 1 contains predictive methods under the training-split estimated null prior. No method dominates all four metrics: verifier-weighted SC gives the best ARC AURC, MAP softmax gives the best ARC ECE, and RA-BSC gives the best ARC accuracy and Brier score among the methods in the table. On CommonsenseQA, pooled majority and classical calibration remain competitive, while RA-BSC improves accuracy and Brier over Bayesian softmax but not AURC. On StrategyQA, Bayesian softmax has the best ECE and Brier, whereas RA-BSC has higher accuracy and lower AURC than Bayesian softmax; none of these predictive differences yields a nonempty LTT certificate. The prior is estimated without calibration or test labels.

Table 1: Main predictive results over 20 question-level splits (mean ± standard deviation). Lower ECE, Brier, and AURC are better.

| Dataset | Method | Accuracy | ECE ↓ | Brier ↓ | AURC ↓ |
|---|---|---|---|---|---|
| ARC-Challenge | Pooled majority | 0.866 ± 0.013 | 0.070 ± 0.016 | 0.093 ± 0.009 | 0.040 ± 0.008 |
| ARC-Challenge | Verifier-weighted SC | 0.876 ± 0.013 | 0.077 ± 0.014 | 0.086 ± 0.011 | 0.038 ± 0.008 |
| ARC-Challenge | Generator-confidence weighted SC (CISC-style) | 0.863 ± 0.013 | 0.075 ± 0.015 | 0.097 ± 0.010 | 0.042 ± 0.008 |
| ARC-Challenge | MAP softmax | 0.867 ± 0.018 | 0.066 ± 0.016 | 0.093 ± 0.010 | 0.046 ± 0.013 |
| ARC-Challenge | Bayesian softmax (mean) | 0.866 ± 0.017 | 0.077 ± 0.018 | 0.093 ± 0.010 | 0.046 ± 0.014 |
| ARC-Challenge | RA-BSC (mean) | 0.888 ± 0.017 | 0.059 ± 0.023 | 0.081 ± 0.009 | 0.054 ± 0.030 |
| CommonsenseQA | Pooled majority | 0.807 ± 0.020 | 0.080 ± 0.013 | 0.133 ± 0.013 | 0.074 ± 0.013 |
| CommonsenseQA | Verifier-weighted SC | 0.799 ± 0.019 | 0.117 ± 0.012 | 0.136 ± 0.014 | 0.071 ± 0.013 |
| CommonsenseQA | Generator-confidence weighted SC (CISC-style) | 0.803 ± 0.021 | 0.085 ± 0.015 | 0.137 ± 0.013 | 0.074 ± 0.013 |
| CommonsenseQA | MAP softmax | 0.792 ± 0.019 | 0.064 ± 0.020 | 0.129 ± 0.011 | 0.085 ± 0.017 |
| CommonsenseQA | Bayesian softmax (mean) | 0.792 ± 0.020 | 0.064 ± 0.016 | 0.130 ± 0.010 | 0.085 ± 0.017 |
| CommonsenseQA | RA-BSC (mean) | 0.818 ± 0.019 | 0.070 ± 0.020 | 0.123 ± 0.012 | 0.107 ± 0.065 |
| StrategyQA | Pooled majority | 0.634 ± 0.027 | 0.229 ± 0.023 | 0.277 ± 0.018 | 0.252 ± 0.025 |
| StrategyQA | Verifier-weighted SC | 0.644 ± 0.026 | 0.265 ± 0.024 | 0.297 ± 0.021 | 0.248 ± 0.026 |
| StrategyQA | Generator-confidence weighted SC (CISC-style) | 0.645 ± 0.028 | 0.236 ± 0.022 | 0.278 ± 0.018 | 0.253 ± 0.024 |
| StrategyQA | MAP softmax | 0.642 ± 0.031 | 0.077 ± 0.027 | 0.225 ± 0.013 | 0.272 ± 0.042 |
| StrategyQA | Bayesian softmax (mean) | 0.641 ± 0.032 | 0.075 ± 0.023 | 0.224 ± 0.012 | 0.270 ± 0.042 |
| StrategyQA | RA-BSC (mean) | 0.662 ± 0.029 | 0.122 ± 0.039 | 0.234 ± 0.019 | 0.254 ± 0.040 |

Table 2: Controlled model decomposition using the same evidence. Deterministic fusion is an unconstrained learned comparator; fixed reliability removes question dependence; adaptive MAP adds question dependence without posterior integration; RA-BSC integrates over the adaptive reliability parameters.

| Dataset | Model | Accuracy | Brier ↓ | AURC ↓ |
|---|---|---|---|---|
| ARC-Challenge | Deterministic fusion | 0.878 | 0.100 | 0.083 |
| ARC-Challenge | Fixed reliability | 0.894 | 0.080 | 0.060 |
| ARC-Challenge | Adaptive MAP | 0.889 | 0.084 | 0.061 |
| ARC-Challenge | RA-BSC (mean) | 0.888 | 0.081 | 0.054 |
| CommonsenseQA | Deterministic fusion | 0.802 | 0.130 | 0.093 |
| CommonsenseQA | Fixed reliability | 0.820 | 0.123 | 0.066 |
| CommonsenseQA | Adaptive MAP | 0.822 | 0.125 | 0.094 |
| CommonsenseQA | RA-BSC (mean) | 0.818 | 0.123 | 0.107 |
| StrategyQA | Deterministic fusion | 0.639 | 0.232 | 0.287 |
| StrategyQA | Fixed reliability | 0.665 | 0.227 | 0.238 |
| StrategyQA | Adaptive MAP | 0.673 | 0.232 | 0.253 |
| StrategyQA | RA-BSC (mean) | 0.662 | 0.234 | 0.254 |

### 5.2 Posterior uncertainty and selective decisions

Scalar calibrators can reduce ECE directly and are not dominated by evidence fusion. We report the Bayesian-softmax and RA-BSC scores separately: the former is the baseline posterior, while the latter marginalizes the question-dependent source-reliability layer. For both models, posterior means are predictive confidences and lower quantiles are selective scores only. Table 9 reports fixed 10% and 5% lower-quantile sensitivities for Bayesian softmax, whereas RA-BSC uses its prespecified 10% lower quantile; none is selected from test outcomes.

Table 2 separates modeling flexibility from Bayesian integration. Relative to Bayesian softmax, RA-BSC improves ARC accuracy from $0.866$ to $0.888$ and Brier score from $0.093$ to $0.081$, and CommonsenseQA accuracy from $0.792$ to $0.818$ and Brier score from $0.130$ to $0.123$. The controlled staircase is not monotone: fixed reliability has the best ARC accuracy and Brier score, adaptive MAP has the best CommonsenseQA and StrategyQA accuracy, and both MAP variants outperform posterior marginalization in StrategyQA accuracy. The unconstrained deterministic fusion comparator is weaker than the structured reliability models. The evidence therefore attributes the point-prediction

Table 3: Mean posterior source reliability. These are diagnostic trust estimates under the model, not observed source accuracies.

| Dataset | 1.5B | 3B | 7B | Answer verifier | Process verifier |
|---|---|---|---|---|---|
| ARC-Challenge | 0.469 | 0.666 | 0.778 | 0.556 | 0.408 |
| CommonsenseQA | 0.375 | 0.498 | 0.622 | 0.441 | 0.326 |
| StrategyQA | 0.219 | 0.253 | 0.316 | 0.221 | 0.222 |

Table 4: Candidate recall by split. The training recall determines the smoothed null prior; other splits are descriptive audits.

| Dataset | Split | Candidate recall | Miss rate |
|---|---|---|---|
| ARC-Challenge | train | $0.953 \pm 0.024$ | 0.047 |
| ARC-Challenge | calibration | $0.956 \pm 0.016$ | 0.044 |
| ARC-Challenge | test | $0.952 \pm 0.010$ | 0.048 |
| CommonsenseQA | train | $0.909 \pm 0.022$ | 0.091 |
| CommonsenseQA | calibration | $0.916 \pm 0.022$ | 0.084 |
| CommonsenseQA | test | $0.919 \pm 0.015$ | 0.081 |
| StrategyQA | train | $0.919 \pm 0.026$ | 0.081 |
| StrategyQA | calibration | $0.914 \pm 0.018$ | 0.086 |
| StrategyQA | test | $0.932 \pm 0.014$ | 0.068 |

gains primarily to structured source reliability, not to Bayesian marginalization. Marginalization instead provides a distribution over question-dependent trust and the uncertainty-bearing ranking scores evaluated below.

Table 7 applies the identical LTT protocol to fixed reliability, adaptive MAP, and RA-BSC. RA-BSC posterior-mean coverage is $0.882$ on ARC and $0.611$ on CommonsenseQA; its prespecified $10\%$ lower-quantile sensitivity gives $0.898$ and $0.546$, respectively. Additional posterior conservatism therefore helps slightly on ARC but hurts CommonsenseQA. Fixed and adaptive variants remain competitive—fixed reliability reaches $0.632$ coverage on CommonsenseQA—so posterior integration is not a universal replacement for deterministic adaptive weighting. StrategyQA has no nonempty LTT certificate for any variant.

### 5.3 Evidence conflict

To connect the reliability layer to its intended use, we reuse the released generator traces to define a question-level disagreement score: the mean pairwise total variation distance between the three generator vote distributions. Table 5 compares Bayesian softmax with RA-BSC on the top quartile of this score, without new generation or threshold tuning. RA-BSC improves high-conflict accuracy from $0.661$ to $0.739$ on ARC and from $0.545$ to $0.614$ on CommonsenseQA; the smaller StrategyQA change, $0.564$ to $0.594$, is not conclusive under the clustered audit. Because each question occurs in multiple randomized test splits, Table 6 uses a question-clustered bootstrap: it first averages correctness within each question across held-out appearances, then resamples unique questions. Against Bayesian softmax, the top-quartile differences are $+0.073$ (95% CI $[0.018, 0.129]$) on ARC and $+0.067$ ($[0.018, 0.117]$) on CommonsenseQA; StrategyQA is inconclusive ($+0.016$, $[-0.047, 0.078]$). The split-wise quartile staircase in Appendix Table 12 shows that deterministic reliability accounts for most of the high-conflict accuracy gain. RA-BSC is consequently not presented as the best point predictor in every conflict regime. Its additional object is posterior uncertainty over the question-dependent trust function, which can be propagated into selective scores.

### 5.4 Focused cross-family audit

The ARC audit in Table 16 replaces the smallest Qwen generator with Gemma-2B while keeping the protocol fixed. Fixed reliability gives the highest point accuracy ($0.892$), slightly above adaptive MAP ($0.888$) and RA-BSC ($0.887$). The ordering reverses for selective ranking: RA-BSC q10 gives the lowest AURC ($0.035$) and the highest mean LTT coverage ($0.917$), compared with $0.058/0.856$

Table 5: Existing-data high-conflict audit. A single question-level disagreement score is formed by averaging across held-out appearances; the resulting fixed top-25% question set is pooled across appearances. Entries are pooled test accuracy. This estimand differs from the split-wise quartiles in Table 12.

| Dataset | Subset | Bayesian softmax | RA-BSC |
|---|---|---|---|
| ARC-Challenge | Bottom 75% | 0.943 | 0.943 |
| ARC-Challenge | Top 25% | 0.661 | 0.739 |
| CommonsenseQA | Bottom 75% | 0.877 | 0.889 |
| CommonsenseQA | Top 25% | 0.545 | 0.614 |
| StrategyQA | Bottom 75% | 0.669 | 0.687 |
| StrategyQA | Top 25% | 0.564 | 0.594 |

Table 6: Question-clustered high-conflict accuracy differences for RA-BSC minus each baseline. A question can occur in multiple test splits; we first average its correctness across those appearances and then resample unique question clusters.

| Dataset | Baseline | $\Delta$ accuracy | 95% CI | Clusters | Split rows |
|---|---|---|---|---|---|
| ARC-Challenge | Fixed reliability | -0.020 | [-0.034, -0.007] | 84 | 760 |
| ARC-Challenge | Adaptive MAP | -0.006 | [-0.022, +0.008] | 84 | 760 |
| ARC-Challenge | Bayesian softmax (mean) | +0.073 | [+0.018, +0.129] | 84 | 760 |
| CommonsenseQA | Fixed reliability | -0.004 | [-0.026, +0.020] | 141 | 1260 |
| CommonsenseQA | Adaptive MAP | -0.015 | [-0.043, +0.011] | 141 | 1260 |
| CommonsenseQA | Bayesian softmax (mean) | +0.067 | [+0.018, +0.117] | 141 | 1260 |
| StrategyQA | Fixed reliability | -0.010 | [-0.024, +0.003] | 143 | 1260 |
| StrategyQA | Adaptive MAP | -0.035 | [-0.060, -0.010] | 143 | 1260 |
| StrategyQA | Bayesian softmax (mean) | +0.016 | [-0.047, +0.078] | 143 | 1260 |

for fixed reliability and $0.065/0.804$ for adaptive MAP. Thus the cross-family mixture reproduces the paper's central separation rather than universal Bayesian dominance: structured reliability drives strong aggregation, while posterior marginalization is most visible in uncertainty-aware selection.

### 5.5 Source reliability carries interpretable uncertainty

Table 3 shows the new posterior object. The inferred ordering varies by benchmark: the 7B generator has the highest mean reliability on all three tasks, while answer and process verifiers exchange relative position. These values are model-implied trust weights, not estimates of $P(\text{source is correct})$ obtained from an independently labeled source-audit set. To quantify uncertainty about source trust itself, we define $U_i^{\text{src}} = |\mathcal{J}|^{-1} \sum_{j \in \mathcal{J}} \text{SD}_q[r_j(E_i; \boldsymbol{\theta}_j)]$ and compare error rates in its lower and upper quartiles in Table 11. This is a diagnostic of whether uncertain source weighting tracks errors, not a training target or a selection rule. The association is strongest on ARC, modest on CommonsenseQA, and reversed in the aggregate StrategyQA diagnostic; consequently, the table supports question-dependent trust modeling but not a universal claim that source uncertainty alone predicts failure.

### 5.6 Why can a lower posterior quantile change certified coverage?

For predicted answer $i$, define the RA-BSC answer-posterior width $U_i^{(.10)} = \bar{p}_{\text{RA},i} - q_{.10,\text{RA},i}$. This is distinct from source-trust uncertainty $U_i^{\text{src}}$ in Table 11: the former measures uncertainty in the candidate posterior, whereas the latter averages posterior dispersion in $r_j(E_i; \theta_j)$. A lower quantile would be immaterial to LTT if it were a monotone transform of the corresponding posterior mean. Table 10 shows that the RA-BSC q10 score is not: the rank correlation is below one, and the composition of the top-ranked half changes by the reported percentage. The high-$U^{(.10)}$ quartile is more error-concentrated than the low-$U^{(.10)}$ quartile. Posterior uncertainty therefore carries additional risk-ranking information, and moderate conservatism can improve certifiability, whereas

overly conservative quantiles need not help. This analysis is purely diagnostic and is not used for model selection or threshold tuning.

### 5.7 RISK CONTROL

Across the evaluated test splits, Table 7 reports nonempty certificates, realized test-risk exceedances, and mean coverage separately for Bayesian softmax and RA-BSC scores. These are empirical audits of finite test splits, not replacements for the LTT guarantee. For ARC, RA-BSC mean and q10 coverage are $0.882$ and $0.898$; for CommonsenseQA they are $0.611$ and $0.546$. No reported method, including either RA-BSC score, produces a nonempty StrategyQA certificate. This reflects both lower base accuracy and limited LTT power: only 30% of 500 questions are available for calibration while 21 thresholds receive Bonferroni control, so the result should not be read as pure prediction failure. The candidate-recall audit explains why the null prior is data-driven: mean training recall is $0.953$ on ARC, $0.909$ on CommonsenseQA, and $0.919$ on StrategyQA, with all estimates formed without calibration or test labels.

Table 7: LTT audit at target risk $0.20$. Test-risk exceedances are empirical diagnostics.

| Dataset | Method | Nonempty certificates / 20 | Test risk exceedances / 20 | Mean coverage |
|---|---|---|---|---|
| ARC-Challenge | Bayesian softmax (mean) | 19 | 0 | 0.835 |
| ARC-Challenge | Fixed reliability | 18 | 0 | 0.854 |
| ARC-Challenge | Adaptive MAP | 19 | 0 | 0.885 |
| ARC-Challenge | RA-BSC (mean) | 19 | 0 | 0.882 |
| ARC-Challenge | RA-BSC (10% lower) | 19 | 0 | 0.898 |
| CommonsenseQA | Bayesian softmax (mean) | 16 | 0 | 0.565 |
| CommonsenseQA | Fixed reliability | 16 | 0 | 0.632 |
| CommonsenseQA | Adaptive MAP | 15 | 0 | 0.607 |
| CommonsenseQA | RA-BSC (mean) | 15 | 0 | 0.611 |
| CommonsenseQA | RA-BSC (10% lower) | 13 | 0 | 0.546 |
| StrategyQA | Bayesian softmax (mean) | 0 | 0 | 0.000 |
| StrategyQA | Fixed reliability | 0 | 0 | 0.000 |
| StrategyQA | Adaptive MAP | 0 | 0 | 0.000 |
| StrategyQA | RA-BSC (mean) | 0 | 0 | 0.000 |
| StrategyQA | RA-BSC (10% lower) | 0 | 0 | 0.000 |

## 6 LIMITATIONS

RA-BSC requires auxiliary verifiers and closed-answer extraction. Its fractional-count likelihood is a conditional evidence model, not a generative account of correlated sources; the resulting generalized-posterior dispersion is not calibrated frequentist uncertainty, and q10 is only a ranking score. Source-duplication and Laplace audits show that dependence and covariance approximation can affect uncertainty rankings (Tables 14 and 13), while inferred reliabilities remain model-implied trust weights. Posterior scores neither replace LTT validity nor universally improve ranking or coverage. Finally, the CISC-style oracle-information diagnostic and focused ARC Qwen–Gemma audit do not establish broad cross-task or cross-family generality.

## 7 CONCLUSION

RA-BSC models candidate answers together with question-dependent evidence-source trust. Across the main and Qwen–Gemma audits, structured reliability improves aggregation under conflict, while fixed or adaptive MAP can outperform posterior marginalization for point prediction. Posterior scores instead provide selective information: q10 helps ARC but hurts CommonsenseQA coverage, and StrategyQA remains uncertifiable. Ranking, calibration, and certifiable abstention are therefore distinct objectives; Bayesian marginalization supplies uncertainty-aware ranking, while LTT supplies finite-sample risk control.

### AI USE STATEMENT

Generative AI tools assisted with writing and polishing; retrieval and discovery of related work; research ideation and execution, including code scaffolding and debugging; and drafting sections of the paper. They were not treated as experimental evidence. The author reviewed the implementation, protocol, numerical artifacts, citations, and manuscript.

### ETHICS STATEMENT

This work uses public benchmarks and generated text without personal data. Selective abstention is not a substitute for expert review in high-stakes settings. The LTT statement relies on exchangeability between calibration and test questions.

### REPRODUCIBILITY STATEMENT

The artifact contains analysis code, generation and verifier prompt constructors, question-level splits, output hashes, per-question predictions, source-reliability summaries, and scripts that reproduce the main reported tables from the released CSV outputs.

## REFERENCES


Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2022.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Gemma 2 Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

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/.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, 2017.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*, 2019.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. doi: 10.18653/v1/2025.findings-acl.1030.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations*, 2025.

Qingni Wang, Yue Fan, and Xin Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025.

# A Additional Results and Statistical Artifacts

The released artifact retains per-question predictions, paired bootstrap summaries, source-reliability posterior summaries, and label-swap audits. These materials provide implementation and uncertainty diagnostics beyond the main comparison.

## A.1 Implementation Details

RA-BSC uses prior precision $\lambda = 1$ and source-likelihood strength 1. Probabilities entering Eq. 7 are clipped below at $\epsilon = 10^{-8}$. MAP parameters are initialized at zero and optimized with L-BFGS-B for at most 300 iterations, using SciPy's convergence criteria (there is no learning-rate hyperparameter). The observed Hessian is computed by central differences with relative step $10^{-4}$; eigenvalues are floored at $10^{-6}$. Unless an audit states otherwise, predictive integrals and q10 scores use 100 Laplace draws per split and seed. The released configuration files and analysis code record these settings exactly.

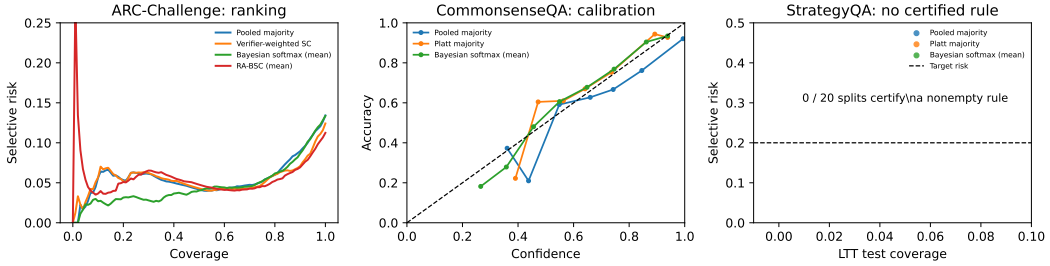


Figure 2: Three observed regimes: verifier-weighted voting is a strong ARC ranking baseline; classical calibration remains competitive on CommonsenseQA; no method certifies a nonempty StrategyQA rule at the stated LTT target.

Table 8: Classical calibration of majority-vote confidence. Platt and beta maps are fitted on the 20% training partition only; the disjoint 30% calibration partition is used exclusively to choose LTT thresholds.

| Dataset | Method | ECE ↓ | Brier ↓ |
|---|---|---|---|
| ARC-Challenge | Pooled majority | $0.070 \pm 0.016$ | $0.093 \pm 0.009$ |
| ARC-Challenge | Platt majority | $0.069 \pm 0.024$ | $0.102 \pm 0.012$ |
| ARC-Challenge | Beta-calibrated majority | $0.080 \pm 0.029$ | $0.103 \pm 0.016$ |
| CommonsenseQA | Pooled majority | $0.080 \pm 0.013$ | $0.133 \pm 0.013$ |
| CommonsenseQA | Platt majority | $0.062 \pm 0.030$ | $0.137 \pm 0.010$ |
| CommonsenseQA | Beta-calibrated majority | $0.072 \pm 0.029$ | $0.136 \pm 0.011$ |
| StrategyQA | Pooled majority | $0.229 \pm 0.023$ | $0.277 \pm 0.018$ |
| StrategyQA | Platt majority | $0.071 \pm 0.038$ | $0.229 \pm 0.008$ |
| StrategyQA | Beta-calibrated majority | $0.086 \pm 0.035$ | $0.232 \pm 0.010$ |

Table 9: Abstention-score ablation. LTT coverage is averaged across splits. Bayesian softmax reports mean, q10, and q5 sensitivities; RA-BSC reports its mean and prespecified q10 score. Lower quantiles affect AURC and LTT ranking only, not predictive calibration.

| Dataset | Abstention score | AURC ↓ | LTT coverage ↑ |
|---|---|---|---|
| ARC-Challenge | Vote fraction | $0.040 \pm 0.008$ | 0.791 |
| ARC-Challenge | MAP softmax | $0.046 \pm 0.013$ | 0.829 |
| ARC-Challenge | Posterior mean | $0.046 \pm 0.014$ | 0.835 |
| ARC-Challenge | Posterior 10% lower | $0.051 \pm 0.019$ | 0.840 |
| ARC-Challenge | Posterior 5% lower | $0.051 \pm 0.020$ | 0.835 |
| ARC-Challenge | RA-BSC posterior mean | $0.054 \pm 0.030$ | 0.882 |
| ARC-Challenge | RA-BSC posterior 10% lower | $0.055 \pm 0.035$ | 0.898 |
| CommonsenseQA | Vote fraction | $0.074 \pm 0.013$ | 0.421 |
| CommonsenseQA | MAP softmax | $0.085 \pm 0.017$ | 0.595 |
| CommonsenseQA | Posterior mean | $0.085 \pm 0.017$ | 0.565 |
| CommonsenseQA | Posterior 10% lower | $0.085 \pm 0.017$ | 0.604 |
| CommonsenseQA | Posterior 5% lower | $0.088 \pm 0.016$ | 0.571 |
| CommonsenseQA | RA-BSC posterior mean | $0.107 \pm 0.065$ | 0.611 |
| CommonsenseQA | RA-BSC posterior 10% lower | $0.091 \pm 0.048$ | 0.546 |
| StrategyQA | Vote fraction | $0.252 \pm 0.025$ | 0.000 |
| StrategyQA | MAP softmax | $0.272 \pm 0.042$ | 0.000 |
| StrategyQA | Posterior mean | $0.270 \pm 0.042$ | 0.000 |
| StrategyQA | Posterior 10% lower | $0.270 \pm 0.039$ | 0.000 |
| StrategyQA | Posterior 5% lower | $0.271 \pm 0.037$ | 0.000 |
| StrategyQA | RA-BSC posterior mean | $0.254 \pm 0.040$ | 0.000 |
| StrategyQA | RA-BSC posterior 10% lower | $0.246 \pm 0.032$ | 0.000 |

Table 10: RA-BSC posterior-quantile mechanism over 20 splits. $U^{(.10)} = \bar{p}_{\mathrm{RA}} - q_{.10,\mathrm{RA}}$; low/high uncertainty errors are bottom/top quartiles.

| Dataset | $\rho(\bar{p}, q_{.10})$ | Top-50% changed | Low-$U^{(.10)}$ error | High-$U^{(.10)}$ error |
|---|---|---|---|---|
| ARC-Challenge | 0.899 | 0.131 | 0.018 | 0.304 |
| CommonsenseQA | 0.935 | 0.104 | 0.099 | 0.427 |
| StrategyQA | 0.975 | 0.101 | 0.241 | 0.419 |

Table 11: Diagnostic source-reliability uncertainty. $U^{\mathrm{src}}$ is the mean posterior SD of question-dependent source reliabilities; quartiles are computed within each test split.

| Dataset | $\rho(U^{\mathrm{src}}, \mathrm{correct})$ | Low-$U^{\mathrm{src}}$ error | High-$U^{\mathrm{src}}$ error |
|---|---|---|---|
| ARC-Challenge | -0.225 | 0.076 | 0.270 |
| CommonsenseQA | -0.088 | 0.166 | 0.252 |
| StrategyQA | 0.073 | 0.366 | 0.272 |

Table 12: Split-wise mean accuracy by conflict quartile, from lowest (Q1) to highest (Q4) disagreement. Quartiles are recomputed within each test split before averaging split-level accuracies; hence Q4 is not the fixed pooled subset in Table 5.

| Dataset | Method | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|
| ARC-Challenge | Fixed reliability | 0.984 | 0.955 | 0.911 | 0.729 |
| ARC-Challenge | Adaptive MAP | 0.984 | 0.955 | 0.911 | 0.712 |
| ARC-Challenge | RA-BSC (mean) | 0.984 | 0.955 | 0.911 | 0.705 |
| CommonsenseQA | Fixed reliability | 0.902 | 0.963 | 0.810 | 0.607 |
| CommonsenseQA | Adaptive MAP | 0.902 | 0.963 | 0.812 | 0.614 |
| CommonsenseQA | RA-BSC (mean) | 0.902 | 0.963 | 0.812 | 0.598 |
| StrategyQA | Fixed reliability | 0.761 | 0.666 | 0.618 | 0.615 |
| StrategyQA | Adaptive MAP | 0.761 | 0.666 | 0.618 | 0.646 |
| StrategyQA | RA-BSC (mean) | 0.761 | 0.666 | 0.618 | 0.603 |

Table 13: Diagonal versus full observed-Hessian Laplace across all 20 splits per benchmark. Differences are full minus diagonal.

| Dataset | Mean $\rho$ | q10 $\rho$ | $\Delta$ Brier | $\Delta$ AURC |
|---|---|---|---|---|
| ARC-Challenge | 0.967 | 0.953 | +0.000 | +0.004 |
| CommonsenseQA | 0.983 | 0.984 | +0.000 | +0.006 |
| StrategyQA | 0.994 | 0.987 | -0.000 | +0.003 |

Table 14: ARC duplicate-source stress test on three representative splits (seeds 0, 7, and 13), averaged across those splits. The duplicated 7B channel contains no new information. These audit values are not the 20-split estimates in the main tables.

| Audit | Original | Duplicated 7B source |
|---|---|---|
| Mean confidence | – | $+0.036$ change |
| Brier | 0.078 | 0.079 |
| AURC | 0.051 | 0.062 |
| LTT coverage | 0.953 | 0.949 |

Table 15: Conservative source-weighting sensitivity across 20 splits. Likelihood strength $0.5$ multiplies every source log-likelihood term in Eq. 7 by $0.5$. Entries are strength-$0.5$ minus standard RA-BSC.

| Dataset | $\Delta$ Accuracy | $\Delta$ Brier | $\Delta$ AURC | $\Delta$ LTT coverage |
|---|---|---|---|---|
| ARC-Challenge | -0.001 | +0.004 | +0.028 | -0.000 |
| CommonsenseQA | -0.001 | +0.005 | -0.032 | +0.039 |
| StrategyQA | -0.006 | -0.006 | -0.018 | +0.000 |

Table 16: Focused cross-family ARC audit over 20 splits. Qwen2.5-1.5B is replaced by Gemma-2-2B-Instruct; Qwen2.5-3B and 7B are retained. Accuracy and Brier use posterior-mean prediction and confidence. The q10 row changes only the selective score used for AURC and LTT.

| Method | Accuracy | Brier ↓ | AURC ↓ | LTT coverage ↑ |
|---|---|---|---|---|
| Bayesian softmax (mean) | 0.868 | 0.095 | 0.049 | 0.850 |
| Verifier-weighted SC | 0.879 | 0.096 | 0.040 | 0.848 |
| Fixed reliability | **0.892** | **0.082** | 0.058 | 0.856 |
| Adaptive MAP | 0.888 | 0.085 | 0.065 | 0.804 |
| RA-BSC (mean) | 0.887 | 0.084 | 0.042 | 0.911 |
| RA-BSC (10% lower) | 0.887 | 0.084 | **0.035** | **0.917** |