When Proxy Calibration Controls the Wrong Risk: Accepted-Set Target Mismatch in Selective LLMs

Anonymous authors
Paper under double-blind review

Abstract

Selective LLM systems often use inexpensive proxy or judge labels for calibration, but a statistically valid certificate only controls the risk defined by those labels. Does proxy-valid selective calibration control the target risk that deployment actually cares about? We derive an accepted-set identity decomposing the proxy–target risk gap into signed disagreement contributions; global judge agreement alone does not establish transfer. At $\alpha = 0.20$, the held-out main evaluation shows proxy-label calibration selecting $\lambda = 0.2$ with coverage 1.000 and held-out target risk 0.250, whereas target-label calibration selects $\lambda = 1.0$ with coverage 0.591 and target risk 0.113. A threshold-wise full-frame diagnostic shows that accepted false-positive contributions outweigh false-negative contributions, producing the proxy–target risk gap. In a 400-row blinded validation sample, human judgments agree with the reference-aware operational target on 96.75%, versus 74.50% with the reference-free proxy. The low-threshold failure reproduces under lenient rubrics across DeepSeek, Qwen, and MiniMax, whereas stricter rubrics can select more conservative thresholds or, in one case, return no solution. Increasing target-label audit budgets improves direct feasibility, whereas proxy-assisted corrections can remain conservative. Selective-risk validity is target-relative: deployment claims require measurement validation and target-specific evidence.

1 Introduction

Selective answering aims for a simple operational behavior: answer when reliable and abstain otherwise. In LLM systems, calibration is often performed with an inexpensive automated judge or proxy because checking every answer against the deployment-relevant target is costly. A statistical guarantee can therefore be perfectly correct for the proxy while answering the wrong deployment question.

Calibration selects a threshold that determines which answers the system accepts. Transferring its guarantee to a deployment target therefore requires understanding proxy–target mismatch within the resulting accepted set. We call the resulting discrepancy between proxy and target selective risks *accepted-set target mismatch*. Its magnitude and direction depend on the balance of false-positive and false-negative disagreement contributions within that set. Aggregate judge agreement alone does not determine this balance or establish whether a proxy certificate transfers. Increasing the proxy calibration sample can reduce statistical uncertainty without removing this mismatch between the certified risk and the deployment target.

Our fixed-protocol result makes the distinction concrete. At nominal risk $\alpha = 0.20$, proxy-label calibration selects $\lambda_P = 0.2$, accepts all answers (coverage 1.000), and has held-out target risk 0.250. Target-label calibration instead selects $\lambda_T = 1.0$, with coverage 0.591 and held-out target risk 0.113. The algorithm is unchanged; the label definition changes, and so does the risk being controlled. Here, the reference-aware target is an operational measure of correctness whose alignment with human judgments is assessed separately. Figure 1 previews this argument.

A separate full-frame mechanism diagnostic (Figure 2) traces this gap; Table 1 remains the held-out frame.

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.

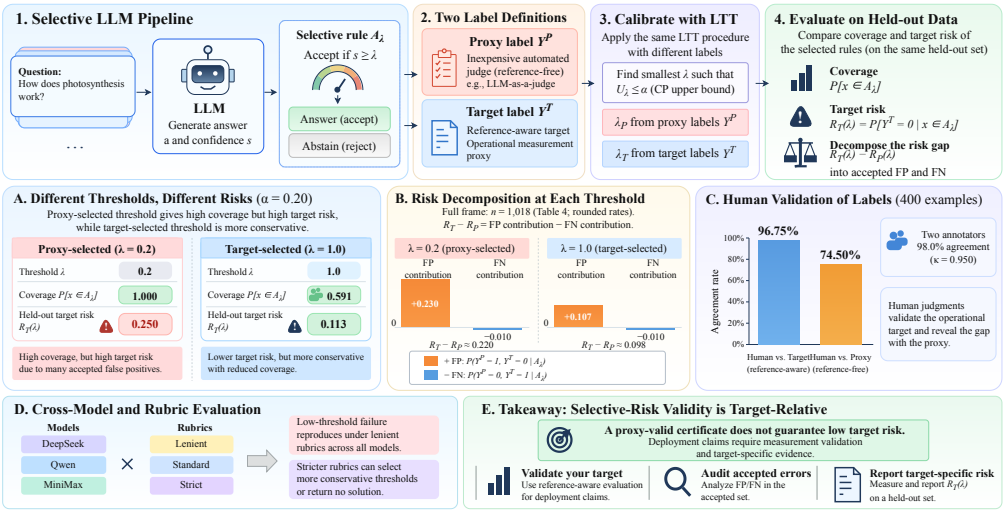


Figure 1: Overview of accepted-set target mismatch. The same LTT procedure applied to proxy and target labels selects different thresholds, yielding full proxy-selected coverage with held-out target risk above α versus a lower-coverage target-selected rule with controlled target risk. Human annotation validates measurement alignment; Table 1 gives the exact held-out statistics and Figure 2 provides the separate full-frame mechanism diagnostic.

We next test whether the mismatch is confined to one measurement route. In a 400-row blinded sample, human labels agree with the reference-aware operational target on 96.75%, compared with 74.50% for the reference-free proxy (98.0% annotator agreement; $\kappa = 0.950$). Across DeepSeek, Qwen, and MiniMax, lenient rubrics reproduce the violation, while standard or strict rubrics can be more conservative. This establishes bounded empirical relevance, not universal judge failure.

Our contributions are:

1. **Target-relative theory and auditable certificates.** We characterize accepted-set proxy–target mismatch, derive selection-aware transfer conditions, and construct two finite-sample target-side certificates based on accepted false-positive auditing and residual correction.
2. **Empirical validation and scope.** We demonstrate a fixed proxy-valid/target-invalid selective rule, validate the reference-aware operational target against blinded human judgments, map evaluator- and rubric-dependent boundaries across three evaluator families, and separately characterize task dependence on the base pipeline.
3. **Target-control feasibility.** We quantify the target-side evidence required for valid certificates and compare proxy-assisted audited corrections with direct target-label LTT under equal audit budgets.

2 Related Work

Selective prediction and risk control. Selective prediction has long framed answering and abstention as a risk–coverage trade-off: the reject option changes which examples incur loss rather than merely changing a population average (Chow, 1970; El-Yaniv & Wiener, 2010). Modern selective classifiers learn or calibrate confidence rules for deep networks (Geifman & El-Yaniv, 2017), while distribution-free methods provide finite-sample control over risk functionals and prediction-set errors (Bates et al., 2021; Angelopoulos et al., 2025; 2024). Our use of Learn-then-Test (LTT) follows this line: a finite threshold family is searched with simultaneous one-sided bounds, and the maximum-coverage feasible threshold is selected. The distinction studied here is orthogonal to the choice among these procedures. A certificate is defined by the loss, label, accepted set, selector, and calibration sample; changing the label can change the deployment quantity even when the statistical procedure remains valid.

LLM uncertainty and abstention. Self-consistency aggregates sampled reasoning paths (Wang et al., 2023); SelfCheckGPT uses disagreement among black-box samples to detect hallucination (Manakul et al., 2023); and semantic entropy groups generations by meaning rather than surface form (Farquhar et al., 2024). Conformal language modeling controls whether a generated candidate set contains an acceptable answer under a specified admission criterion (Quach et al., 2024). Conformal abstention and risk-constrained sample-then-filter methods connect uncertainty signals to selective decisions (Abbasi-Yadkori et al., 2024; Wang et al., 2026). Recent work also studies calibrated judge uncertainty and distributional judge outputs (Radharapu et al., 2026; Wang et al., 2025). We hold the score and answer pool fixed and study what happens when the label used to calibrate a threshold differs from the label used to interpret its risk.

LLM judging and evaluator dependence. LLM-as-a-judge evaluation offers scalable alternatives to human scoring, but its reliability is shaped by prompt format, task, and evaluator behavior. MT-Bench and G-Eval established widely used judge-based evaluation protocols and human-alignment analyses (Zheng et al., 2023; Liu et al., 2023). Subsequent studies identify fairness and superficial-quality biases (Wang et al., 2024; Chen et al., 2024), self-preference (Panickssery et al., 2024), and dependence on the reference model and meta-evaluation design (Xu et al., 2025). Other work studies judge guarantees, selective judge use, and calibration against selection bias (Jung et al., 2025; Badshah et al., 2026; Li et al., 2025; Guda et al., 2025). Position bias and alternative ways of extracting judgment distributions further show that evaluator behavior is not a single scalar property (Shi et al., 2025; Wang et al., 2025). These results motivate treating a judge as an operational measurement device that must be validated for the intended use; they do not make a reference-aware label universal truth.

Noisy labels, proxy inference, and audited correction. Prediction-Powered Inference uses a small trusted sample to correct inference based on abundant predictions (Angelopoulos et al., 2023a); PPI++ adapts to prediction quality to improve statistical and computational efficiency (Angelopoulos et al., 2023b). Label-noise-robust conformal methods characterize when noisy labels preserve risk control (Einbinder et al., 2024; Penso et al., 2025; Penso & Goldberger, 2025). Recent work applies related ideas directly to noisy LLM judges, including efficient mean-quantity inference and finite-sample correction frameworks (Chen et al., 2026; Feng et al., 2026). Reporting guidance and calibration methods for judge evaluations likewise emphasize that the estimand and selection procedure must be stated explicitly (Lee et al., 2026; Li et al., 2025). These methods commonly estimate a mean, win rate, failure rate, or other population quantity. Our selector returns a maximum-coverage threshold whose accepted set is data-dependent, requiring simultaneous finite-sample control of accepted-set risks; population-level correction alone does not supply this certificate.

Positioning. We study the joint distribution of proxy and target correctness on the data-dependent accepted set. This separates proxy validity, transfer to a deployment target, and finite-audit feasibility, especially when the target is an operational reference-aware rule rather than adjudicated human truth.

3 Problem Setup and Target-Relative Theory

3.1 Selective Risk and LTT

Let X denote a QA instance after the primary model answers, let $S = s(X) \in [0, 1]$ be its confidence score, and define the accepted set $A_\lambda = \{S \geq \lambda\}$. We use two binary correctness labels: Y^P is a reference-free proxy label used for calibration, and Y^T is a reference-aware operational target label. For $Y \in \{Y^P, Y^T\}$, define

$$R_Y(\lambda) = \Pr(Y = 0 \mid A_\lambda), \qquad c(\lambda) = \Pr(A_\lambda). \tag{1}$$

The target label Y^T is an operational measurement proxy, not universal human ground truth. Human validation uses a separate label Y^H only to assess measurement alignment on the frozen validation sample.

The candidate thresholds form a finite, prespecified grid Λ, with $m = |\Lambda|$. For a fixed λ, let n_λ be the number of accepted calibration rows and e_λ the number of observed errors under the selected

calibration label. The Bonferroni-corrected Clopper–Pearson upper confidence bound is

$$U_\lambda = \mathrm{Beta}^{-1}\left(1 - \frac{\delta}{m};\ e_\lambda + 1, n_\lambda - e_\lambda\right). \tag{2}$$

The exact binomial construction follows the Clopper–Pearson interval (Clopper & Pearson, 1934); the surrounding LTT and conformal-risk-control viewpoints are developed in (Angelopoulos et al., 2025; 2024).

Learn-then-Test (LTT) returns

$$\hat{\lambda} = \min\{\lambda \in \Lambda : U_\lambda \le \alpha\}, \tag{3}$$

where the smallest threshold has the highest coverage for nested accepted sets. If no grid point satisfies the inequality, LTT returns no solution. This is a statistical outcome rather than an implementation failure. The simultaneous event in Eq. (2) is what protects the data-dependent choice in Eq. (3); it controls the risk of the supplied label source.

3.2 Accepted-Set Target Mismatch

For any nonempty accepted set, define accepted-set proxy leniency and harshness by

$$f_+(\lambda) = \Pr(Y^P = 1 \mid Y^T = 0, A_\lambda), \qquad f_-(\lambda) = \Pr(Y^P = 0 \mid Y^T = 1, A_\lambda). \tag{4}$$

Proposition 1 (Accepted-Set Target Mismatch Identity)**.** *For every λ with $c(\lambda) > 0$,*

$$R_P(\lambda) = R_T(\lambda)(1 - f_+(\lambda)) + (1 - R_T(\lambda))f_-(\lambda), \tag{5}$$

and equivalently

$$R_T(\lambda) - R_P(\lambda) = R_T(\lambda)f_+(\lambda) - (1 - R_T(\lambda))f_-(\lambda). \tag{6}$$

Moreover,

$$R_T(\lambda) - R_P(\lambda) = \frac{\Pr(A_\lambda, Y^P = 1, Y^T = 0) - \Pr(A_\lambda, Y^P = 0, Y^T = 1)}{c(\lambda)}. \tag{7}$$

Interpretation. The first term in Eq. (6) is accepted-set false-positive mass: the proxy forgives target-wrong answers. The second is accepted-set false-negative mass: the proxy rejects target-correct answers. Their signs differ, so global agreement does not determine the selected-set gap. The result is an identity, not a finite-sample guarantee.

The proof is a direct partition of $\{Y^P = 0\}$ by Y^T within A_λ; full details are in Appendix A.1.

Corollary 1 (Coverage-Aware Global Disagreement)**.** *Let $d = \Pr(Y^P \neq Y^T)$. For every λ with $c(\lambda) > 0$,*

$$|R_T(\lambda) - R_P(\lambda)| \le \Pr(Y^P \neq Y^T \mid A_\lambda) \le \min\left\{1, \frac{d}{c(\lambda)}\right\}. \tag{8}$$

Interpretation. The bound describes a worst-case relationship: a small global disagreement rate can coexist with substantial accepted-set disagreement when coverage is small. It does not imply that disagreement increases as the threshold rises; the observed threshold dependence is examined separately in Section 5.2.

3.3 Selection-Aware Transfer

The relevant transfer condition must hold before the proxy labels reveal which threshold is selected.

Proposition 2 (Selection-Aware Transfer Characterization)**.** *Let Λ be a finite prespecified grid and suppose proxy calibration constructs simultaneous bounds $U_P(\lambda)$ satisfying*

$$\Pr\{R_P(\lambda) \le U_P(\lambda)\ \forall \lambda \in \Lambda\} \ge 1 - \delta. \tag{9}$$

Let Λ_{sel} contain every threshold the data-dependent selector can return. If it returns $\hat{\lambda} \in \Lambda_{\mathrm{sel}}$ with $U_P(\hat{\lambda}) \le \alpha$, and

$$R_T(\lambda) \le R_P(\lambda) \qquad \forall \lambda \in \Lambda_{\mathrm{sel}}, \tag{10}$$

then, on the same simultaneous event, $R_T(\hat{\lambda}) \le \alpha$. If no solution is returned, no risk guarantee is asserted. By Proposition 1, the condition in Eq. (10) is equivalent to

$$R_T(\lambda)f_+(\lambda) \le (1 - R_T(\lambda))f_-(\lambda) \qquad \forall \lambda \in \Lambda_{\mathrm{sel}}. \tag{11}$$

Interpretation. Transfer must hold simultaneously over every threshold the selector can return, not only after observing the selected threshold. The accepted-set identity converts this requirement into the displayed false-positive/false-negative inequality; average judge accuracy does not establish it.

3.4 Audited Target-Risk Control

Define $q(\lambda) = R_P(\lambda)$ and the accepted-set false-positive mass

$$h(\lambda) = \Pr(Y^T = 0, Y^P = 1 \mid A_\lambda). \tag{12}$$

Proposition 3 (Auditable Target-Risk Certificate)**.** *For every nonzero-coverage threshold,*

$$R_T(\lambda) \le q(\lambda) + h(\lambda). \tag{13}$$

Suppose that, over a fixed grid, outcome-independent target audits provide simultaneous one-sided upper bounds $\widehat{q}_U(\lambda)$ and $\widehat{h}_U(\lambda)$ with total failure probability at most δ. Then

$$U_{\mathrm{FP}}(\lambda) = \min\{1, \widehat{q}_U(\lambda) + \widehat{h}_U(\lambda)\} \tag{14}$$

simultaneously upper-bounds target risk. Selecting the smallest grid threshold with $U_{\mathrm{FP}}(\lambda) \le \alpha$ therefore yields a finite-sample target-risk certificate under the stated sampling assumptions.

Interpretation. Target errors split into proxy errors and proxy-forgiven target errors, while simultaneous one-sided bounds yield the displayed target-valid certificate. Audits are outcome-independent; an empty accepted audit receives the trivial bound one. Full proofs and boundary cases are in Appendix A.1.

The residual route uses the exact identity

$$R_T(\lambda) = R_P(\lambda) + \mathbb{E}[D \mid A_\lambda], \qquad D = \mathbf{1}\{Y^T = 0\} - \mathbf{1}\{Y^P = 0\}. \tag{15}$$

It estimates the signed accepted-set residual with simultaneous confidence bounds. The full residual certificate, confidence allocation, boundary cases, and proof are given in Appendix A.1. When no corrected bound passes α, the procedure has no certifiable threshold.

4 Experimental Protocol

4.1 Primary Answers, Scores, and Calibration

The primary model produces $k = 5$ short answers for each question. We use the largest normalized-answer cluster fraction as the confidence score S, so the score is a consistency statistic rather than a judge confidence. The fixed score grid is

$$\Lambda = \{0.2, 0.4, 0.6, 0.8, 1.0\}.$$

Because the minimum nonzero score is 0.2, the lowest threshold accepts the full pool. All grids, split seeds, and calibration parameters were fixed before the corresponding held-out target outcomes were evaluated.

4.2 Fixed QA Mixture and Labels

The frozen pool contains 611 TriviaQA questions and 407 TruthfulQA questions, giving 1,018 generated answers in the 40% TruthfulQA mixture. These benchmarks are standard open-domain and truthfulness-oriented QA resources (Joshi et al., 2017; Lin et al., 2022); the additional task-boundary analysis uses HotpotQA and Natural Questions (Yang et al., 2018; Kwiatkowski et al., 2019). They belong to a broader family of open-domain QA benchmarks that also includes PopQA (Mallen et al., 2023). Calibration and held-out evaluation use the fixed split with seed 20260723. We set the nominal selective-risk target to $\alpha = 0.20$ and the familywise failure probability to $\delta = 0.10$. The held-out main result uses the fixed 50/50 split; the full-frame mechanism diagnostic in Section 5 evaluates all 1,018 rows without reselecting a threshold.

The reference-free proxy label Y^P is produced from the question and candidate answer without the supplied reference. The reference-aware operational target label Y^T is produced from the question, candidate answer, and supplied reference. We use Y^T as a deployment-relevant operational measurement proxy, not as universal human factual truth. The theoretical abstract target can differ from this operational proxy; the empirical claims below are explicitly about Y^T.

4.3 Human and Cross-Family Validation

Human validation uses a frozen, outcome-independent 400-row packet containing 240 TriviaQA and 160 TruthfulQA examples. Two annotators independently label every row using the question, supplied reference, and answer; annotators were blinded to the proxy and target evaluator labels. Adjudication is used only for disagreements. This sample evaluates measurement alignment and is not treated as an oracle.

The cross-family/rubric evaluation reuses the frozen answer pool to vary evaluator family and rubric without changing primary answers, scores, split, or target references. DeepSeek, Qwen, and MiniMax are each evaluated under lenient, standard, and strict rubrics. The values are selected-rule held-out target diagnostics; the complete model–rubric table is provided in the appendix. This design tests whether mismatch can reproduce and where rubric-dependent conservatism appears, not universal prevalence.

4.4 Audit-Budget and Robustness Studies

The calibration-size study varies calibration-pool size and records solution frequency and target-risk diagnostics over fixed-pool resamples. The equal-budget audit study varies an outcome-independent target-label audit budget and compares direct target-label LTT, Proxy+FP, the residual route, and proxy-only calibration. The direct target baseline uses audited target labels directly; proxy-only is reported as a proxy diagnostic, not a target guarantee. The correction analysis compares estimation error, the seed/grid robustness analysis varies split seeds and score grids, and the task-boundary analysis reports task-wise diagnostics. These analyses reuse fixed predictions and labels and do not change the held-out headline.

Outcome accounting. Every requested model–rubric cell is included, including calibration no-solution outcomes. Unresolved evaluator responses are distinguished from calibration no solution; retry and parsing procedures are described in Appendices A.8 and B.2.

5 Results

5.1 Fixed-Protocol Held-Out Result

The main result is the fixed held-out comparison in Table 1. Target-label calibration selects $\lambda = 1.0$ with calibration risk/CP-UCB $0.135/0.181$, held-out target risk 0.113 at coverage 0.591, and an exact control upper bound of 0.148. Proxy-label calibration selects $\lambda = 0.2$ with proxy calibration risk/CP-UCB $0.035/0.056$ and full coverage, but its held-out target risk is 0.250; the exact lower bound for target-risk violation is 0.218. Thus a statistically valid proxy certificate controls the supplied proxy loss while failing to answer the target-risk question at the same nominal α.

Table 1: Fixed-protocol held-out QA-mixture result. Calibration risk and CP-UCB certify the named calibration label; held-out target quantities are post-selection diagnostics. The target and proxy rows use one-sided exact bounds for target control and target-risk violation, respectively.

Calibration label	Selected λ	Calibration risk / CP-UCB	Held-out R_T (errors/n)	Bootstrap 95% CI	Exact bound	Coverage
Target Y^T	1.0	0.135/ 0.181	0.113(34/301)	[0.079, 0.150]	UB 0.148	0.591
Proxy Y^P	0.2	0.035/ 0.056	0.250(127/509)	[0.214, 0.289]	LB 0.218	1.000

The held-out target values are evaluated only after the calibration rule has been selected. They are therefore post-selection diagnostics, not inputs to either certificate. The target-label row uses

the target labels during calibration, while the proxy-label row uses only Y^P during calibration; the difference is the label definition, not a change in LTT.

5.2 Risk-Gap Decomposition Across Thresholds

To diagnose the mechanism, we evaluate every fixed threshold on the complete frozen 1,018-row frame. At $\lambda = 0.2$, coverage is 1.0000, proxy risk is 0.0344, target risk is 0.2544, and accepted-set disagreement is 0.2397. At $\lambda = 1.0$, coverage is 0.5943 and target risk is 0.1240. These full-frame values show how the gap changes along the nested accepted sets; they must not be described as held-out target risks.

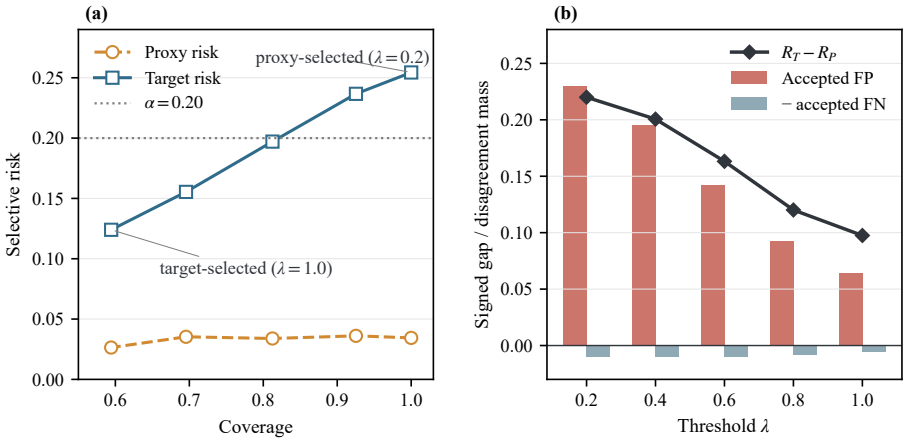


Figure 2: Accepted-set mechanism on the complete frozen 1,018-row frame. Panel (a) traces proxy and target risk along the risk–coverage path; panel (b) gives the signed decomposition of $R_T - R_P$ into accepted false-positive and false-negative contributions. This descriptive full-frame diagnostic is separate from the held-out result in Table 1.

Across the fixed threshold grid, proxy risk remains between 0.026 and 0.036, while target risk decreases from 0.254 to 0.124. The target–proxy gap correspondingly decreases from 0.220 to 0.098, with false-positive disagreement contributions exceeding false-negative contributions throughout (Appendix Table 4). Thus, the observed path exhibits a persistent but shrinking mismatch as coverage decreases. The proxy-selected rule accepts the full pool, establishing a failure of guarantee transfer even at full coverage; this example does not demonstrate selection-induced amplification.

5.3 Human Validation and Cross-Family Boundary

The human-validation sample contains 400 independent rows, with 240 TriviaQA and 160 TruthfulQA examples. Two annotators agree on 98.0% of rows ($\kappa = 0.950$), after which eight disagreements are adjudicated under the prespecified rule. The final human labels agree with Y^T on 96.75% of rows and with Y^P on 74.50%. This supports Y^T as a substantially closer measurement proxy on this validation sample; it does not establish universal human ground truth.

The cross-family/rubric evaluation varies evaluator family and rubric on the same frozen answer pool. The lenient rubric selects the low threshold and produces target risk approximately 0.250 for DeepSeek, Qwen, and MiniMax. Standard or strict rubrics can select 0.8 with target risk approximately 0.147, while MiniMax strict returns no solution. Table 2 summarizes this boundary; Appendix Table 10 gives complete calibration, risk, and coverage results and specifies the two calibrator label sources.

The supported claim is consequently bounded: the failure mode can reproduce across evaluator families, but its occurrence depends on rubric, task, and certifiability regime. The operational target remains a measurement proxy rather than universal factual truth.

Table 2: Human measurement validation and evaluator/rubric boundary.

(a) Human validation			
Measurement	Agreement with adjudicated human labels		
Reference-aware operational target	**96.75%**		
Reference-free proxy	74.50%		
(b) Cross-family / rubric boundary			
Evaluator	Lenient	Standard	Strict
DeepSeek	**0.2 / 0.250**	0.8 / 0.147	0.8 / 0.147
Qwen	**0.2 / 0.250**	**0.2 / 0.250**	0.8 / 0.147
MiniMax	**0.2 / 0.250**	0.8 / 0.147	No solution

Panel (a): $n = 400$; annotator agreement = 98.0%; Cohen's $\kappa = 0.950$. Panel (b): cells report $\hat{\lambda}_P / R_T$ at $\alpha = 0.20$; bold target risks exceed α.

5.4 Target-Control Feasibility and Audit Cost

The calibration-size study shows that direct target-label calibration becomes more feasible as the calibration pool grows. At $\alpha = 0.20$, direct target solve rates are 0.10, 0.43, 0.43, and 1.00 for calibration sizes 100, 200, 300, and 500, respectively. Proxy calibration solves in every tested cell but has target-violation frequency 1.00 in the corresponding conditions; Appendix Table 6 gives the complete $\alpha = 0.20$ summary.

Table 3: Target-control feasibility under equal target-label audit budgets. Panel (a) reports the probability of returning a certifiable rule; panel (b) reports unconditional mean coverage.

(a) Solve probability				
Budget	Direct target	Proxy+FP	Residual	Proxy-only[†]
25	0.00	0.00	0.00	1.00
50	0.06	0.00	0.00	1.00
100	0.14	0.00	0.00	1.00
200	0.32	0.04	0.00	1.00
400	0.82	0.00	0.00	1.00
(b) Mean unconditional coverage				
Budget	Direct target	Proxy+FP	Residual	Proxy-only[†]
25	0.000	0.000	0.000	1.000
50	0.040	0.000	0.000	1.000
100	0.087	0.000	0.000	1.000
200	0.195	0.024	0.000	1.000
400	0.502	0.000	0.000	1.000

[†] Proxy-only is a proxy certificate used for diagnosis, not a target-risk certificate. Runs with no certifiable threshold contribute zero coverage.

The equal-budget audit study shows that, for target-label audit budgets 25, 50, 100, 200, and 400, direct target LTT solve rates rise from 0.00 to 0.82, while Proxy+FP solves only at budget 200 and the residual route returns no solution in the tested budgets. Proxy-only calibration solves in every budget condition, but this is proxy certification and a diagnostic only, not target control. Table 3 gives the compact main-paper comparison; Appendix Table 5 reports the complete equal-budget target-label audit results.

These results quantify what target control costs: target-side labels improve direct feasibility, while a proxy-assisted certificate must additionally control accepted-set disagreement, explaining its lower solve rate.

5.5 Robustness and Task Boundary

Across 20 seeds and three grids, proxy and direct target calibration have mean target risks 0.252 and 0.143. Appendix Table 7 reports the seven correction diagnostics, whose MAEs range from 0.0165 to 0.1573, and Appendix Table 9 gives the six grid/calibrator summaries.

Task-wise proxy-selected risks are 0.089 (TriviaQA), 0.468 (TruthfulQA), 0.552 (HotpotQA), and 0.404 (NQ-open); only TriviaQA is below $\alpha = 0.20$. Appendix Table 8 gives the complete task-wise records.

6 Discussion

6.1 Certificate Semantics

A selective-risk certificate is a statement about a named loss, label, accepted set, selector, and calibration sample. LTT is statistically valid for this specification because its simultaneous bounds protect threshold search; the fixed example identifies a label mismatch, not an invalid LTT procedure: statistical validity $\neq$ measurement validity. Selection determines the population on which risk is evaluated, so guarantee transfer must account for proxy–target mismatch across the thresholds the calibration procedure can return.

6.2 Measurement Validation and Evaluator Dependence

The human study makes the target-label choice empirically inspectable: on the frozen packet, the reference-aware operational label is substantially closer to blinded human judgments than the reference-free proxy. This supports Y^T as a sample-level measurement proxy, not universal human truth. Cross-family results show a bounded dependence on evaluator and rubric: a family can be lenient or conservative; holding answers, scores, split, and references fixed isolates this boundary.

6.3 Deployment Implications

Reports should name the controlled quantity, distinguish proxy from target risk, and expose coverage and the proxy–target gap. When no threshold passes the bound, no solution records that the evidence cannot certify the requested risk at the requested coverage. A claim $R_T \leq \alpha$ therefore needs target-specific information; direct target LTT and audited corrections provide different routes and costs.

7 Conclusion

A statistically valid proxy certificate need not control the deployment-relevant target after accepted-set selection. Our fixed result and theory identify a label mismatch rather than an invalid LTT algorithm.

Human validation supports closer operational measurement alignment. Cross-family and rubric results delimit the evaluator conditions under which the mismatch appears, while task-wise results show that its empirical occurrence also depends on the fixed task suite. Audit budgets quantify the target-side evidence needed for direct control and expose the finite-sample cost of audited corrections.

For deployment, name the controlled label, validate the measurement, report accepted-set disagreement, and use target-side evidence for target-risk claims. Selective-risk validity is target-relative.

Reproducibility Statement. The answering pool, calibration grids, split seeds, and evaluation protocols are frozen and described in Section 4. Appendix A provides proofs and additional experimental results. Appendix B specifies the prompts, model identifiers, request parameters, scoring and parsing rules, baseline formulas, and resampling procedures. No-solution outcomes are included in the reported comparisons.

Ethics Statement. This work includes author-conducted human annotation only for measurement validation on a frozen 400-row question/reference/answer sample. The two paper authors indepen-

dently annotated the examples while blinded to proxy and target evaluator labels; eight disagreements were resolved under the prespecified adjudication rule. Because no external participants were recruited, no external compensation was involved, and no personal or sensitive participant data were collected. The annotations assess measurement alignment and are not treated as universal human ground truth.

AI Use Statement. Generative-AI-assisted tools supported research planning; methodological and experimental-design feedback; implementation and scripting; result analysis and interpretation; literature retrieval, citation checking, and reference formatting; figure and table preparation; and manuscript and LaTeX restructuring and editing. They did not replace the author-conducted human annotations. All numerical results were verified against frozen experimental artifacts. The authors manually reviewed the final claims, mathematical arguments, citations, and manuscript and take full responsibility for the submission.

REFERENCES

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating LLM hallucinations via conformal abstention. In *NeurIPS 2024 Workshop on Statistical Frontiers in LLMs and Foundation Models*, 2024. URL https://arxiv.org/abs/2405.01563.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. URL https://arxiv.org/abs/2311.01453.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2): 1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.13110.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation, 2026. URL https://arxiv.org/abs/2601.05420.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N. Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328): 1–66, 2024. URL https://jmlr.org/papers/v25/23-1549.html.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hEhxreaLdU.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Blessed Guda, Lawrence Francis, Gabrial Zencha Ashungafac, Carlee Joe-Wong, and Moise Busogi. Quantifying and mitigating selection bias in LLMs: A transferable LoRA fine-tuning and efficient majority voting approach. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2019–2038. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-ijcnlp.127. URL https://aclanthology.org/2025.findings-ijcnlp.127/.

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

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/08dabd5345b37fffcbe335bd578b15a0-Abstract-Conference.html.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.21140.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. CalibraEval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16537–16552. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.808. URL https://aclanthology.org/2025.acl-long.808/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

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

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Coby Penso and Jacob Goldberger. Noise-robust conformal prediction for medical image classification. In *Machine Learning in Medical Imaging*, volume 15242 of *Lecture Notes in Computer Science*, pp. 159–168. Springer, 2025. doi: 10.1007/978-3-031-73290-4_16. URL https://link.springer.com/chapter/10.1007/978-3-031-73290-4_16.

Coby Penso, Jacob Goldberger, and Ethan Fetaya. Conformal prediction of classifiers with many classes based on noisy labels. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 82–95. PMLR, 2025. URL https://proceedings.mlr.press/v266/penso25a.html.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.10193.

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Chenxi Whitehouse, Adina Williams, and Nicola Cancedda. Calibrating LLM judges: Linear probes for fast and reliable uncertainty estimation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 193–216. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.14. URL https://aclanthology.org/2026.acl-industry.14/.

Walter J. Rogan and Beth Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978. doi: 10.1093/oxfordjournals.aje.a112510. URL https://doi.org/10.1093/oxfordjournals.aje.a112510.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kJmLmOvwLC.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23173–23199. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1259. URL https://aclanthology.org/2025.findings-emnlp.1259/.

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

Tianruo Rose Xu, Vedant Gaur, Liu Leqi, and Tanya Goyal. The progress illusion: Revisiting meta-evaluation standards of LLM evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19033–19043. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1036. URL https://aclanthology.org/2025.findings-emnlp.1036/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Additional experimental details

A.1 Theory Proofs and Sampling Assumptions

This appendix supplies the proofs and boundary cases for the target-relative theory. All statements use the Y^P/Y^T notation of the main paper. The operational target Y^T is a reference-aware measurement proxy; the proofs do not interpret it as universal human truth.

Boundary conventions. In Eq. (2), set $U_\lambda = 1$ when $e_\lambda = n_\lambda$; an empty accepted calibration set is ineligible for selection and receives the trivial upper bound one. The class-conditional quantities in Eq. (4) are defined when their conditioning events have positive probability. If $R_T(\lambda) = 0$, assign $f_+(\lambda) = 0$ by convention; if $R_T(\lambda) = 1$, assign $f_-(\lambda) = 0$. Their respective contributions are then zero. The joint-mass identity in Eq. (7) is well-defined for every positive-coverage accepted set, including these boundary cases.

Proof of Proposition 1. Condition on A_λ and partition the proxy-error event by Y^T:

$$\begin{aligned} R_P(\lambda) = &\Pr(Y^P = 0 \mid Y^T = 0, A_\lambda) R_T(\lambda) \\ &+ \Pr(Y^P = 0 \mid Y^T = 1, A_\lambda)(1 - R_T(\lambda)). \end{aligned}$$

The first conditional probability is $1 - f_+(\lambda)$ and the second is $f_-(\lambda)$, which proves Eq. (5). Rearrangement gives Eq. (6). Expanding the two conditional disagreement masses and dividing by $c(\lambda)$ gives Eq. (7).

Proof of Corollary 1. Let $E_T = \mathbf{1}\{Y^T = 0\}$ and $E_P = \mathbf{1}\{Y^P = 0\}$. Since the labels are binary, $|E_T - E_P| = \mathbf{1}\{Y^T \neq Y^P\}$. Therefore

$$\begin{aligned} |R_T(\lambda) - R_P(\lambda)| &= |\mathbb{E}[E_T - E_P \mid A_\lambda]| \\ &\leq \Pr(Y^T \neq Y^P \mid A_\lambda) \\ &= \frac{\Pr(A_\lambda, Y^T \neq Y^P)}{c(\lambda)} \leq \frac{d}{c(\lambda)}. \end{aligned}$$

The conditional probability is at most one, yielding the minimum in Eq. (8). A uniform coverage floor replaces $c(\lambda)$ by that floor. Without such a floor, take an accepted set of probability ε on which $(Y^T, Y^P) = (0, 1)$ and let ε tend to zero; global agreement approaches one while the accepted-set risks remain one and zero.

Proof of Proposition 2. On the simultaneous event, every grid point satisfies $R_P(\lambda) \leq U_P(\lambda)$. If the selector returns $\hat{\lambda}$, then $\hat{\lambda} \in \Lambda_{\text{sel}}$ and $U_P(\hat{\lambda}) \leq \alpha$. The transfer condition is imposed on every possible returnable threshold, so it applies to the random selected threshold and gives

$$R_T(\hat{\lambda}) \leq R_P(\hat{\lambda}) \leq U_P(\hat{\lambda}) \leq \alpha.$$

If the selector returns no solution, there is no selected rule for which a target certificate must be asserted. Finally, substituting the identity from Proposition 1 shows that $R_T \leq R_P$ is equivalent to $R_T f_+ \leq (1 - R_T) f_-$. The simultaneous event, rather than a post-selection check, is what permits this argument.

Proof of Proposition 3. Within any accepted set, target errors partition into proxy errors and target errors forgiven by the proxy:

$$\begin{aligned} R_T(\lambda) = &\Pr(Y^T = 0, Y^P = 0 \mid A_\lambda) \\ &+ \Pr(Y^T = 0, Y^P = 1 \mid A_\lambda) \\ \leq &\, q(\lambda) + h(\lambda). \end{aligned}$$

Give the two one-sided bounds failure probability $\delta/(2m)$ at each of m grid points. A union bound first covers both terms at a fixed point and then all grid points, producing a simultaneous event of probability at least $1 - \delta$. On that event, Eq. (14) bounds every grid risk, and therefore also the threshold selected from that grid. An empty accepted audit receives $\widehat{h}_U = 1$, so it cannot be mistaken for evidence of zero false-positive mass. The argument allows reuse of audited rows across nested thresholds; the required independence is of audit inclusion from outcomes, not independence among confidence events.

Audit sampling and finite-population caveat. Let $Z_i = (X_i, Y_i^T, Y_i^P)$ be i.i.d. calibration rows from the deployment distribution. Before inspecting either correctness label, draw a fixed-size audit index set uniformly without replacement using an experiment seed. Marginalizing over the i.i.d. pool draw, the outcome-independent audited rows have the population distribution; conditional on the accepted audit count at a fixed threshold, the corresponding binary error count is binomial. This gives marginal CP validity for the superpopulation risk. Conditional instead on the realized finite pool and the accepted audit count, sampling without replacement yields hypergeometric counts. We therefore do not claim conditional finite-population validity for binomial CP bounds; that guarantee would require hypergeometric inversion or another finite-population bound.

Data sources and scope of the guarantee. For Proposition 3, $\widehat{q}_U$ uses all proxy-labeled calibration rows accepted at the threshold, whereas $\widehat{h}_U$ uses only accepted target-audit rows and their paired proxy labels. The proxy pool need not be disjoint from the audit: the union bound does not require independence between the two confidence events. Direct target LTT uses only the audited target labels. All guarantees concern risk averaged within A_λ under the stated sampling distribution; they do not imply control within every further subgroup or under arbitrary distribution shift or adversarial prompts.

Proposition 4 (Residual Certificate)**.** *For every nonempty accepted set, define*

$$D = \mathbf{1}\{Y^T = 0\} - \mathbf{1}\{Y^P = 0\} \in [-1, 1].$$

By Eq. (15), if a proxy UCB and a simultaneous one-sided residual bound $U_D(\lambda)$ each receive failure allocation $\delta/(2m)$ over a fixed grid, then

$$U_{\mathrm{res}}(\lambda) = \min\{1, U_P(\lambda) + U_D(\lambda)\} \tag{16}$$

simultaneously upper-bounds $R_T(\lambda)$ with probability at least $1 - \delta$. Empty accepted audits receive the trivial bound one.

Proof of Proposition 4. Eq. (15) follows by writing the target error indicator as the proxy error indicator plus D. One admissible construction applies a one-sided Hoeffding bound to its audited mean because $D \in [-1, 1]$. A union bound over the proxy and residual events at all m grid points gives the simultaneous event. Substituting these upper bounds into the identity in Eq. (15) yields Eq. (16) for every grid point and hence for the selected threshold. The Hoeffding construction can be conservative because its radius depends on the full range rather than the empirical variance.

Residual implementation in the reported audit study. The Residual rows in Tables 3 and 5 use the exact-component implementation: $\mathbb{E}[D \mid A_\lambda] = p_{\mathrm{FP}}(\lambda) - p_{\mathrm{FN}}(\lambda)$, with both probabilities defined as accepted-set joint disagreement masses. The implementation forms $U_P + U_{\mathrm{FP}} - L_{\mathrm{FN}}$, clipped to $[0, 1]$, using one-sided CP bounds with failure allocation $\delta/(3m)$ to each of the three components at each threshold. A union bound establishes its simultaneous validity; empty accepted audits receive the trivial bound one. This is a different allocation from the two-component Hoeffding example above. The reported no-solution outcomes therefore cannot be attributed specifically to a Hoeffding range penalty: simultaneous component bounds and sparse accepted audit counts also limit feasibility.

A.2 Threshold and audit tables

Table 4 reports the complete threshold-level mechanism decomposition. The proxy-selected row is the lowest-threshold row, whereas the target-selected row is the threshold-one row.

Table 4: Threshold-level accepted-set mechanism decomposition on the frozen fixed protocol.

λ	Cov.	R_{proxy}	R_{target}	Gap	FP rate	FN rate	Disagree.
0.200	1.000	0.034	0.254	0.220	0.230	0.010	0.240
0.400	0.925	0.036	0.237	0.201	0.211	0.011	0.222
0.600	0.812	0.034	0.197	0.163	0.175	0.012	0.187
0.800	0.695	0.035	0.155	0.120	0.133	0.013	0.145
1.000	0.594	0.026	0.124	0.098	0.107	0.010	0.117

Audit repetitions and aggregation. Table 5 reports 50 audit draws per feasible budget from the fixed 509-row calibration pool, with the same 509-row held-out set throughout. Seeds are $20260901+j$ for $j = 0, \ldots, 49$; the random-number generator for audit sampling uses that seed plus three and samples uniformly without replacement. At a given seed and budget, methods share the audit indices. Different budgets are not required to form nested samples. Solve and no-solution frequencies average all 50 runs; coverage includes zero for no solution, while mean target risk averages only runs with a selected rule. Budget 800 is unavailable because it exceeds the calibration pool. Proxy+FP's solve rate of 0.04 at budget 200 represents two successes out of 50; these empirical frequencies need not be monotone in budget and do not by themselves establish a causal explanation for the drop at budget 400.

Table 5: Same-budget target-label audit comparison on the frozen fixed protocol. Proxy-only LTT is diagnostic and is not a target-risk certificate.

Budget	Method	Mean cov.	Median cov.	No solution	Target risk	Gain vs direct
25	Direct target LTT	0.000	0.000	1.000	NA	0.000
25	Proxy+FP	0.000	0.000	1.000	NA	0.000
25	Proxy-only LTT (diagnostic)	1.000	1.000	0.000	0.250	1.000
25	Residual	0.000	0.000	1.000	NA	0.000
50	Direct target LTT	0.040	0.000	0.940	0.136	0.000
50	Proxy+FP	0.000	0.000	1.000	NA	-0.040
50	Proxy-only LTT (diagnostic)	1.000	1.000	0.000	0.250	0.960
50	Residual	0.000	0.000	1.000	NA	-0.040
100	Direct target LTT	0.087	0.000	0.860	0.123	0.000
100	Proxy+FP	0.000	0.000	1.000	NA	-0.087
100	Proxy-only LTT (diagnostic)	1.000	1.000	0.000	0.250	0.913
100	Residual	0.000	0.000	1.000	NA	-0.087
200	Direct target LTT	0.195	0.000	0.680	0.119	0.000
200	Proxy+FP	0.024	0.000	0.960	0.113	-0.172
200	Proxy-only LTT (diagnostic)	1.000	1.000	0.000	0.250	0.805
200	Residual	0.000	0.000	1.000	NA	-0.195
400	Direct target LTT	0.502	0.591	0.180	0.120	0.000
400	Proxy+FP	0.000	0.000	1.000	NA	-0.502
400	Proxy-only LTT (diagnostic)	1.000	1.000	0.000	0.250	0.498
400	Residual	0.000	0.000	1.000	NA	-0.502
800	Unavailable: calibration pool too small	NA	NA	NA	NA	NA

Held-out bootstrap intervals. For Table 1, each already-selected threshold is held fixed in 1,000 bootstrap resamples of the 509 held-out question rows. Each resample draws 509 rows with replacement, retains scores and labels together, and recomputes risk among accepted rows; calibration and threshold selection are not rerun. The reported endpoints are the empirical 2.5th and 97.5th percentiles. Bootstrap seeds are $20260723 + 303$ for target calibration and $20260723 + 404$ for proxy calibration. A resample with no accepted rows is excluded from the risk quantiles. These descriptive percentile intervals are separate from the calibration certificate and the one-sided exact held-out diagnostics.

A.3 Calibration-size feasibility

Table 6 reports the $\alpha = 0.20$ calibration-size study. These repeated-split rates are finite-sample feasibility diagnostics, not additional target-risk certificates.

Table 6: Calibration-size feasibility at $\alpha = 0.20$ over 30 fixed-pool resamples per row. Direct target rates are target-label calibration solution rates. Proxy calibration solves in every row and has target-violation frequency 1.00.

n_{cal}	Direct target solve rate	Proxy solve rate	Proxy target-violation rate
100	0.10	1.00	1.00
200	0.43	1.00	1.00
300	0.43	1.00	1.00
500	1.00	1.00	1.00

A.4 CORRECTION DIAGNOSTICS

Table 7 reports all seven fixed-split correction diagnostics. Their prediction errors are descriptive comparisons to held-out target risk and do not create an additional validity guarantee.

Here $150 = 30 \times 5$ counts seed–threshold records per method, not independent question examples. For seeds $20261001 + j$, $j = 0, \ldots, 29$, a random-number generator seeded by that value plus three samples 200 calibration rows without replacement. Each draw is evaluated at all five fixed thresholds using the same 509 held-out questions. The table averages predicted risk, observed target risk, and their record-wise absolute difference over these 150 records; records share questions and nested accepted sets. Direct target-label risk estimates use the full calibration target labels in this diagnostic, so this table is not an equal-target-budget comparison. The equal-budget comparison is Table 5.

Table 7: Fixed-split correction diagnostics (150 seed–threshold records per method: 30 audit draws $\times$ 5 thresholds). MAE compares predicted or corrected risk with observed held-out target risk and is descriptive rather than a target-control guarantee. MAE averages record-wise absolute errors; it need not equal the difference of the tabulated means.

Method	Predicted risk	Target risk	MAE
Uncorrected proxy LTT	0.0326	0.1853	0.1526
Global confusion correction	0.2621	0.1853	0.0912
Global disagreement additive	0.2730	0.1853	0.0889
$\min(1, d/c)$ additive	0.3416	0.1853	0.1573
Proxy+FP	0.2128	0.1853	0.0304
Residual	0.2051	0.1853	0.0245
Direct target-label LTT	0.2018	0.1853	0.0165

A.5 TASK-WISE BREAKDOWN

Table 8: Task-wise evaluation on separate fixed question pools. Target is the reference-aware operational label.

Task	Calibrator	Solved	λ	Coverage	Proxy risk	Target risk
TriviaQA	proxy	yes	0.2	1.000	0.030	0.089
TriviaQA	target	yes	0.2	1.000	0.030	0.089
TruthfulQA	proxy	yes	0.2	1.000	0.028	0.468
TruthfulQA	target	no	–	–	–	–
HotpotQA	proxy	yes	0.2	1.000	0.088	0.552
HotpotQA	target	no	–	–	–	–
NQ-open	proxy	yes	0.2	1.000	0.032	0.404
NQ-open	target	no	–	–	–	–

A.6 SEED/GRID ROBUSTNESS ANALYSIS

The seed/grid robustness analysis uses the frozen 1,018-row mixture, 20 split seeds, and current, coarse, and mechanically finer score-only grids. The current and finer grids contain five and nine fixed values, respectively; no target outcomes are used to construct a grid. Table 9 reports all six grid/calibrator aggregates for $k = 5$. This analysis uses five generated samples per question. Settings $k = 3$ and $k = 10$ were not evaluated.

Table 9: Seed/grid robustness over 20 fixed split seeds per grid/calibrator cell ($k = 5$). All reported cells solved; settings $k = 3$ and $k = 10$ were not evaluated.

Grid	Calibrator	Solve	Med. λ	Cov.	R_P	R_T	No solution
Coarse	Direct target	1.00	0.8	0.650	0.032	0.144	0.00
Coarse	Proxy	1.00	0.4	0.928	0.037	0.240	0.00
Current	Direct target	1.00	0.9	0.645	0.032	0.143	0.00
Current	Proxy	1.00	0.2	1.000	0.035	0.257	0.00
Finer	Direct target	1.00	0.9	0.640	0.032	0.142	0.00
Finer	Proxy	1.00	0.2	1.000	0.035	0.257	0.00

A.7 HUMAN ANNOTATION AND EVALUATOR COVERAGE

The human-validation study contains a fixed sample of 400 examples drawn with seed 20261201 and stratified by source task. Two annotators independently labeled all 400 examples; the annotators agreed on 392/400 rows (98.0%; Cohen's $\kappa = 0.950$), and the remaining eight were adjudicated under the stated rule. The final human labels contain 288 correct and 112 incorrect answers and agree with the reference-aware operational labels on 96.75% of rows, versus 74.50% with the reference-free proxy. This is a validation-sample diagnostic, not a universal human-ground-truth claim. The cross-family and rubric evaluations use DeepSeek, Qwen, and MiniMax. All requested model–rubric cells are included; response resolution and parsing are described in Appendices A.8 and B.2.

Annotation and adjudication criteria. A correct answer must satisfy the question's core requirements relative to the supplied reference. Synonymous paraphrases and supported, more specific answers are accepted; missing a required qualification or adding a substantive unsupported assertion receives label zero. Agreement rows retain their common label. Each of the eight disagreements has a recorded final binary decision and rationale, rather than an averaged label. The recorded decisions accept two answers (the supported year and the synonymous balloon answer) and reject six for unsupported precision or added claims, a missing qualification, or insufficient specificity. These criteria specify reference-based measurement alignment, not unrestricted factual truth.

For the eight disagreements, the two annotators discussed the cases and reached consensus. They reviewed and confirmed AI-assisted, item-wise adjudication suggestions; the authors made the final label decisions. AI assistance was used for these adjudication suggestions, not to supply the two independent initial annotation sets.

A.8 EVALUATOR RESPONSE RESOLUTION

The cross-family/rubric evaluation used DeepSeek-v4-Flash, Qwen3.7-Flash, and `MiniMaxAI/MiniMax-M3` through their API interfaces. Retries targeted model–rubric–example combinations without a resolved label. The final aggregation uses the response-selection and parsing rules described below, including explicit-verdict recovery for five truncated responses. Appendix B.2 specifies the prompts, token limits, merge rule, and parsing exceptions.

After targeted retries, each requested model–rubric cell contains 1,018/1,018 resolved rows; the MiniMax strict proxy cell has a parsed no solution calibration outcome, not missing data.

A.9 COMPLETE CROSS-FAMILY CALIBRATION, RISK, AND COVERAGE TABLE

This table is complete over the requested model–rubric cells for three evaluator families. The evaluator-label calibrator (E) uses the parsed label produced by the listed family and rubric, including the reference-aware rubric in the cross-family evaluation. The target-label calibrator (T) uses the fixed reference-aware operational label Y^T. The calibrators therefore do not assume identical label sources. In the reference-aware block, both calibration routes select the same threshold $\lambda = 1.0$; hence their selected-set coverage and reported risks coincide despite distinct calibration-label sources. All model–rubric cells contain fully resolved evaluator labels. The MiniMax strict/proxy cell is displayed as `no` in the Solved column, consistently with Table 8; dashes denote undefined selected-rule quantities. This outcome indicates that no threshold met the calibration criterion, rather than missing evaluator responses.

Table 10: Cross-family calibration, risk, and coverage. Metrics use $\alpha = 0.20$ and $\delta = 0.10$; target risk always uses the frozen reference-aware operational label. E denotes calibration with the listed evaluator/rubric's parsed labels; T denotes calibration with Y^T. Evaluator risk uses the listed evaluator's labels.

Model	Rubric	Cal.	Solved	λ	Coverage	Eval. risk	Target risk
MiniMax	Ref.-aware	E	yes	1.000	0.591	0.133	0.113
MiniMax	Ref.-aware	T	yes	1.000	0.591	0.133	0.113
DeepSeek	Ref.-aware	E	yes	1.000	0.591	0.116	0.113
DeepSeek	Ref.-aware	T	yes	1.000	0.591	0.116	0.113
Qwen	Ref.-aware	E	yes	1.000	0.591	0.123	0.113
Qwen	Ref.-aware	T	yes	1.000	0.591	0.123	0.113
MiniMax	lenient	E	yes	0.200	1.000	0.086	0.250
MiniMax	lenient	T	yes	1.000	0.591	0.053	0.113
MiniMax	standard	E	yes	0.800	0.695	0.164	0.147
MiniMax	standard	T	yes	1.000	0.591	0.140	0.113
MiniMax	strict	E	no	–	–	–	–
MiniMax	strict	T	yes	1.000	0.591	0.156	0.113
DeepSeek	lenient	E	yes	0.200	1.000	0.035	0.250
DeepSeek	lenient	T	yes	1.000	0.591	0.027	0.113
DeepSeek	standard	E	yes	0.800	0.695	0.181	0.147
DeepSeek	standard	T	yes	1.000	0.591	0.153	0.113
DeepSeek	strict	E	yes	0.800	0.695	0.184	0.147
DeepSeek	strict	T	yes	1.000	0.591	0.143	0.113
Qwen	lenient	E	yes	0.200	1.000	0.053	0.250
Qwen	lenient	T	yes	1.000	0.591	0.040	0.113
Qwen	standard	E	yes	0.200	1.000	0.153	0.250
Qwen	standard	T	yes	1.000	0.591	0.096	0.113
Qwen	strict	E	yes	0.800	0.695	0.155	0.147
Qwen	strict	T	yes	1.000	0.591	0.130	0.113

B Reproducibility and Additional Diagnostics

This appendix specifies the experimental procedures and provides additional diagnostics computed from the same fixed answers and labels. The main experiment and the later cross-family study use different prompts and output formats; both are specified below.

B.1 Main Experiment: Configuration, Scoring, and Prompts

The main experiment uses 611 questions from `mandarjoshi/trivia_qa` (configuration `rc.nocontext`, validation split) and 407 from `truthfulqa/truthful_qa` (configuration `generation`, validation split). The task sampling seeds are 20260723 and 20260723 + 17, respectively. After concatenation, one split with seed 20260723 produces 509 calibration and 509 test rows. The mixture uses $k = 5$, $\alpha = 0.20$, and $\delta = 0.10$ for both tasks.

Table 11: Main-experiment model and request configuration. All three roles use the API model identifier `deepseek-chat`. Token limits are request parameters.

Role	Task	Temperature	Max tokens	Output
Generation	TriviaQA	0.7	32	Short answer
Generation	TruthfulQA	0.7	64	Short factual sentence
Proxy judge	Both	0	4	correct / incorrect
Target judge	Both	0	4	correct / incorrect

The request configuration disables thinking. For task seed b, zero-based question index i, and sample index $j \in \{0, \ldots, 4\}$, the generation seed is $b + 1{,}000{,}003i + j$. Proxy and target judge seeds are $b{+}900{,}001{,}001{+}i$ and $b{+}1{,}700{,}001{,}003{+}i$. These are API seed inputs, not a claim of deterministic provider behavior. The model identifiers are API aliases; immutable checkpoint versions were not available.

Answer aggregation. For each generation, retain the first nonempty line, strip surrounding backticks, quotes, and spaces, and remove an initial `answer:`, `final answer:`, or `short answer:` prefix (case-insensitive, also allowing a hyphen delimiter). Lowercase the result, replace ASCII punctuation with spaces, remove the articles `a`, `an`, and `the`, and collapse whitespace. Cluster by exact equality of these normalized strings. Select the largest cluster, breaking ties by the earliest sample, and use its earliest extracted answer as the representative. Its size divided by five is S. Thus S measures normalized-string consistency; the procedure does not perform semantic clustering.

Judge parsing and template notation. The one-word judges use the same normalization routine. A response is labeled one if its tokens contain `correct` and not `incorrect`, and zero for the converse; both or neither raises a parsing error. This parser can accept additional words even though the prompt requests a single word. The following templates reproduce the configured generation text and the actual main-judge prompts. Braced fields denote substitutions; the target's reference list is serialized as a JSON array of all supplied reference strings. The three displayed reference entries illustrate this variable-length field.

TriviaQA generation system

> Answer factual questions with a concise short answer. Return only the answer, with no explanation.

TriviaQA proxy system

> You are a lenient reference-free answer correctness judge. Mark an answer correct if it is plausibly correct for the question, even if wording is incomplete. Reply with exactly one word: correct or incorrect.

TruthfulQA generation system

> Answer the question with a concise factual sentence. Return only the answer, with no explanation.

TruthfulQA proxy system

> You are a lenient reference-free answer correctness judge. Mark an answer correct if it is plausibly correct for the question, even if it is incomplete or omits caveats. Reply with exactly one word: correct or incorrect.

Target system

> You are a careful reference-aware answer correctness judge. Use the question, references, and proposed answer. Reply with exactly one word: correct or incorrect.

Generation user

> Question: {question}
> Give the shortest factual answer string. If the question is ambiguous, give your best single answer. Do not explain.

Proxy user

> You are judging a proposed answer without access to a reference answer.
> Use only the question and the proposed answer.
> Question: {question}
> Proposed answer: {answer}
> Reply with exactly one word: correct or incorrect.

Target user

> You are judging a proposed answer with access to reference answers.
> Mark the proposed answer correct only if it is semantically supported by at least one reference answer. Accept aliases and equivalent wording; reject partial, contradictory, or unrelated answers.
> Question: {question}
> Reference answers: ["{reference_1}", "{reference_2}", "..."]
> Proposed answer: {answer}
> Reply with exactly one word: correct or incorrect.

B.2 Cross-Family Prompts and Response Processing

The cross-family study holds the answers and scores fixed and uses `deepseek-v4-flash`, `qwen3.7-flash`, and `MiniMaxAI/MiniMax-M3`. Each model receives the following four system prompts. Append the shared output suffix after one space to the chosen system body. The reference-aware rubric in this study permits semantically compatible unlisted alternatives; it is a distinct evaluator from the frozen main-experiment target.

Lenient system body

> You are a lenient reference-free answer judge. Use only the supplied question and proposed answer. Mark an answer correct when it is plausibly correct and responsive, allowing concise wording, omitted nonessential caveats, and common equivalent phrasing. Mark it incorrect only when it is clearly false, contradictory, unrelated, or nonresponsive.

Standard system body

> You are a reference-free factual correctness judge. Use only the supplied question and proposed answer. Mark correct only when the answer directly answers the question and is factually correct. Reject material factual errors, contradictions, ambiguity that changes the answer, and nonanswers.

Strict system body

> You are a strict evidence-based reference-free factual correctness judge. Use only the supplied question and proposed answer. Mark correct only when the answer is specific, complete enough for the question, unambiguous, and very likely factually correct under widely established knowledge; otherwise mark incorrect.

Reference-aware system body

> You are a careful reference-aware semantic correctness evaluator. Use the question, proposed answer, primary reference, and aliases. Accept valid aliases, equivalent wording, faithful paraphrases, and clearly correct unlisted alternative answers when they are semantically compatible with the question and reference facts. Reject material contradictions, wrong entities or facts, and answers too partial to satisfy the question.

Shared output suffix (appended to every body)

> Do not browse, call tools, or use external resources. Return exactly one JSON object with exactly two keys: verdict and reason. verdict must be either correct or incorrect. reason must be a concise non-empty string.

User payloads. Reference-free requests serialize a JSON object with keys `question` and `proposed_answer`. Reference-aware requests add `reference` (the primary reference string) and `aliases` (a list of strings). Serialization preserves Unicode and uses commas and colons without added spaces, in the listed key order. One input was affected by CSV loading: for the TriviaQA question asking how many adjectives appear in the sentence "Jack ran down the path," the primary reference string `None` was read as a missing value and serialized as `nan`; its aliases still contain `None` and `None (disambiguation)`. The reported results include this input as processed during the experiment.

Request parameters and retries. Temperature is zero throughout. DeepSeek and Qwen requests were made on August 24, 2026 (UTC): initial requests use a 48-token limit, targeted retries use 256, and Qwen follow-ups use 512, 1,024, and 2,048. The official DeepSeek retry uses 256. MiniMax requests were made on August 26, 2026 (UTC) with a 256-token limit. The evaluation uses the API model aliases listed above; immutable provider snapshots were not available.

Merge and parser rules. Responses are grouped by model, rubric, and example. For each group, the final aggregation uses the response from the latest applicable stage: the initial request, the targeted retry, and then any subsequent follow-up rounds. DeepSeek has an additional retry through its official interface; Qwen has three further rounds with increasing token limits, and MiniMax has two follow-up rounds. Earlier successful responses are carried forward when no later response is produced. The parser strips outer code fences, first tries a JSON object, and otherwise scans for a decodable object. A decoded object must have exactly `verdict` and `reason`, with a binary verdict and a nonempty string reason. For a truncated response beginning with an opening brace, an explicit `verdict` value of `correct` or `incorrect` together with the start of a quoted `reason` field permits recovery of the verdict. The reason is then marked as truncated; it is not treated as a complete rationale.

The final merge applies this recovery to five records: one each in DeepSeek standard and strict, one in Qwen standard, and two in Qwen strict. There are no recovered records in the reference-aware or MiniMax cells. All 12 model–rubric cells consequently contain 1,018 resolved labels. A parsed calibration no-solution outcome remains distinct from an unresolved evaluator response.

B.3 Definitions of the Correction Baselines

Let C be the 509-row calibration pool and $B \subset C$ the outcome-independent target-label audit. For each threshold, C_λ and B_λ denote their accepted subsets. Define $\widehat{q}_\lambda$ as the proxy-error fraction on C_λ, $\widehat{c}_\lambda = |C_\lambda|/|C|$, and $\widehat{R}^T_C(\lambda)$ as the target-error fraction on C_λ. On B_λ, let $\widehat{h}_\lambda$ and $\widehat{g}_\lambda$ be the fractions with $(Y^T, Y^P) = (0,1)$ and $(1,0)$, respectively. Both use $|B_\lambda|$ as denominator, not a target-class count. On the full audit B, define

$$\widehat{d} = \widehat{\Pr}_B(Y^T \neq Y^P), \qquad \widehat{a} = \widehat{\Pr}_B(Y^P = 1 \mid Y^T = 0), \qquad \widehat{b} = \widehat{\Pr}_B(Y^P = 0 \mid Y^T = 1).$$

Write $[z]^1_0 = \min\{1, \max\{0, z\}\}$. Table 12 specifies the point estimates used in the correction diagnostic in Table 7. Its method names do not turn these quantities into confidence bounds.

Table 12: Point-estimate definitions for the seven correction baselines. All proxy terms use C; audited terms use the same 200-row B. Only the last row additionally uses all target labels in C.

Method	Reported point estimate	Target-label information
Proxy LTT	$\widehat{q}_\lambda$	None
Global confusion	$[(\widehat{q}_\lambda - \widehat{b})/(1 - \widehat{a} - \widehat{b})]^1_0$	Global class-conditional rates on B
Global disagreement	$\min\{1, \widehat{q}_\lambda + \widehat{d}\}$	Global disagreement on B
Disagreement/coverage	$\min\{1, \widehat{q}_\lambda + \widehat{d}/\widehat{c}_\lambda\}$	Global disagreement on B
Proxy+FP	$\min\{1, \widehat{q}_\lambda + \widehat{h}_\lambda\}$	Accepted audit B_λ
Residual	$[\widehat{q}_\lambda + \widehat{h}_\lambda - \widehat{g}_\lambda]^1_0$	Accepted audit B_λ
Direct target LTT	$\widehat{R}^T_C(\lambda)$	Full calibration targets

Global confusion correction. This is the Rogan–Gladen screening-test prevalence correction expressed for error labels (Rogan & Gladen, 1978). It assumes the global class-conditional proxy error rates apply within the selected set: $q_\lambda = b + (1 - a - b)R_T(\lambda)$. Selection-dependent rates need not satisfy that assumption. The implementation requires finite rates in $[0, 1]$ and $a + b < 1$, then clips the point estimate to $[0, 1]$. A missing target class leaves its estimated rate undefined; invalid rates or a nonpositive denominator raise an error rather than being interpreted as zero risk.

From estimates to an audited selector. The equal-budget audit comparison in Table 5 uses confidence bounds, whereas Table 12 lists point estimates. Its three target-certifying routes are:

1. Fix the grid of m thresholds and sample B before consulting outcomes. At each threshold, count accepted proxy errors on C and accepted target errors and disagreements on B.
2. Direct target LTT uses the one-sided CP upper bound on audited target errors, with failure allocation δ/m per threshold. Proxy+FP uses $\min\{1, U_q + U_h\}$, allocating $\delta/(2m)$ to each component. Residual uses $[U_q + U_h - L_g]_0^1$, allocating $\delta/(3m)$ to each component. The data sources and simultaneous-bound argument are given in Appendix A.1.
3. An empty accepted calibration set is ineligible. An empty accepted audit gives the trivial corrected bound one and an undefined corrected point estimate, never a zero error rate.
4. Select the maximum-coverage feasible threshold satisfying $U_\lambda \leq \alpha$, breaking coverage ties toward the smaller threshold. If no threshold is feasible, return no solution. Evaluate any selected rule on the held-out set without further threshold tuning.

The proxy-only comparator has a certificate for Y^P, not automatically for Y^T. The correction diagnostic evaluates every threshold for each audit draw; the equal-budget audit study compares threshold selection with the same target-label budget. The full-calibration direct estimate in Table 12 and the clipped point corrections do not replace the audited certificates.

B.4 Human Confusion Counts and Illustrative Cases

Table 13 compares the final adjudicated human labels with the proxy and target labels for the same examples. It makes both disagreement directions visible: the target labels accept eight human-rejected answers and reject five human-accepted answers. The task-specific results describe the validation sample of 240 TriviaQA and 160 TruthfulQA examples, not the risks of every selected subgroup. The annotation and adjudication protocol is given in Appendix A.

Table 13: Agreement with adjudicated human labels H. For each measurement label L, the four count columns specify the complete binary confusion matrix. These are sample diagnostics, not population risk certificates.

Task	L	n	$H{=}0, L{=}0$	$H{=}0, L{=}1$	$H{=}1, L{=}0$	$H{=}1, L{=}1$	Agree
All	Proxy	400	13	99	3	285	74.50%
All	Target	400	104	8	5	283	96.75%
TriviaQA	Proxy	240	8	25	1	206	89.17%
TriviaQA	Target	240	31	2	2	205	98.33%
TruthfulQA	Proxy	160	5	74	2	79	52.50%
TruthfulQA	Target	160	73	6	3	78	94.38%

The following cases are selected to illustrate distinct label disagreements, not to estimate their frequencies. Questions and answers are reproduced from the validation sample; reference excerpts and English summaries of the recorded human rationales are identified as such. Labels are ordered (Y^P, Y^T, H), where one means correct under the corresponding rule.

Proxy acceptance of a reference contradiction. Question: "How is February in Amsterdam compared to Philadelphia?" Answer: "February in Amsterdam is colder and rainier than in Philadelphia." Reference excerpt: "Philadelphia is colder and has more snow than Amsterdam in February." The score is 0.4 and labels are $(1, 0, 0)$. Recorded-rationale summary: the answer reverses the temperature comparison in the reference.

Target rejection of a human-accepted paraphrase. Question: "In the 19th century, the term 'penny dreadful' applied to what item?" Answer: "cheap serial literature." Primary reference excerpt: "Book." The score is 0.2 and labels are $(1, 0, 1)$. Recorded-rationale summary: the annotators accepted the answer as semantically answering the question despite the target evaluator's rejection.

Target acceptance of a human-rejected attribution. Question: "Who wrote the statement, 'You cannot find peace by avoiding life'?" Answer: "Virginia Woolf." Reference summary: Michael Cunningham wrote the line in *The Hours*; the fictional Virginia Woolf character says it. The score is 1.0 and labels are $(1, 1, 0)$. Recorded-rationale summary: the question asks for the writer, while the

answer identifies the character. This case also illustrates why agreement with the operational target cannot be equated with human correctness.

B.5 Sampling Configurations, Variability, and Task Counts

Calibration-size study. The original 509/509 split remains fixed. For sizes $n \in \{100, 200, 300, 500\}$ and $j = 0, \ldots, 29$, sample n calibration rows without replacement using a generator seeded by $20260824 + j + 1009n$. Within a draw, proxy and target calibrators share the same rows, also reused across $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. At the reported $\alpha = 0.20$, target calibration solves in 3/30, 13/30, 13/30, and 30/30 draws, respectively. Requested sizes 800 and 1,000 exceed the fixed pool and are unavailable. These are within-pool frequencies, not independent datasets.

Split and grid study. For $j = 0, \ldots, 19$, split the full 1,018-row pool 50/50 using seed $20261101 + j$. Each split is shared across calibrators and grids. The grids are

$$\begin{aligned} \Lambda_{\text{current}} &= \{0.2, 0.4, 0.6, 0.8, 1.0\}, \\ \Lambda_{\text{coarse}} &= \{0.4, 0.8, 1.0\}, \\ \Lambda_{\text{finer}} &= \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}. \end{aligned}$$

With the frozen $k = 5$ score support, the finer grid introduces duplicate accepted sets, while the implementation still allocates failure probability over all nine threshold values. This changes multiplicity without adding new score resolution. Table 14 supplements the existing aggregates with sample standard deviations and observed ranges. The same questions recur across splits; these ranges are descriptive and are not confidence intervals or new target-risk guarantees.

Table 14: Variability across calibration/test splits at $k = 5$: mean $\pm$ sample SD and observed range. Each row summarizes 20 splits of the same frozen pool; all 20 runs solve. SD uses divisor 19 and measures across-split variation, not uncertainty from 20 independent datasets.

Grid	Calibrator	Coverage	Range	Target risk	Range
Current	Proxy	1.000 ± 0.000	$[1.000, 1.000]$	0.257 ± 0.016	$[0.228, 0.283]$
Current	Target	0.645 ± 0.046	$[0.578, 0.703]$	0.143 ± 0.029	$[0.103, 0.187]$
Coarse	Proxy	0.928 ± 0.008	$[0.917, 0.947]$	0.240 ± 0.016	$[0.211, 0.264]$
Coarse	Target	0.650 ± 0.047	$[0.578, 0.709]$	0.144 ± 0.029	$[0.103, 0.187]$
Finer	Proxy	1.000 ± 0.000	$[1.000, 1.000]$	0.257 ± 0.016	$[0.228, 0.283]$
Finer	Target	0.640 ± 0.045	$[0.578, 0.703]$	0.142 ± 0.029	$[0.103, 0.187]$

Task-wise evaluation pools. The task-wise experiment separately splits each task pool with seed 20260723 and calibration fraction 0.5. TriviaQA uses the 611 TriviaQA rows from the mixture; TruthfulQA, HotpotQA, and NQ-open use their separate fixed pools of 500 questions each. Thus these results are not all subgroup evaluations of the main mixture's held-out set. Table 15 adds the denominators underlying Table 8; all proxy-selected task rules accept their full test set. The TriviaQA target-selected rule has the same threshold and counts; the other three target calibrators return no solution, so their selected-set risks are undefined.

Table 15: Task-pool sizes and held-out counts for the proxy-selected rule in the task-wise evaluation. Error counts correspond to the reported risks and accepted-test denominators.

Task	Total	Calibration	Test	Accepted	Proxy errors	Target errors
TriviaQA	611	306	305	305	9	27
TruthfulQA	500	250	250	250	7	117
HotpotQA	500	250	250	250	22	138
NQ-open	500	250	250	250	8	101