FS-CERT-U: RISK CERTIFICATION BEFORE UTILITY OPTIMIZATION FOR NON-MONOTONIC TWO-STAGE MODEL ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-stage model routing trades answer coverage against inference cost, but its system-level selection-conditioned error rate (SCER) can move in either direction as either routing threshold changes. Searching for and certifying policies on the same evidence can also invalidate fixed-policy guarantees. We introduce **FS-Cert-U**, which separates policy construction, ordering, certification, and utility selection: design data freeze a finite policy family, screening data determine its order, independent certification applies exact candidate-wise tests with fixed-sequence stopping, and utility is optimized only within the resulting certified set. We prove finite-sample family-wise safety for the certified set without requiring independence across candidate tests, together with a utility-regret bound relative to the best certified policy. Across 33 QA routing configurations, FS-Cert-U shows a systematic certification-power advantage over matched Bonferroni correction: dataset-level utility gains persist across all six risk targets, and all 33 paired gains are positive at the primary operating point. The advantage also persists across four Guardian-cost weights. Controlled-population experiments directly target false certification and show the predicted decline in post-certification utility regret as certification size grows.

1 INTRODUCTION

Large language models increasingly serve cost-sensitive systems. Routers and cascades answer easy inputs with cheap models and escalate harder cases to stronger models or retrieval, trading inference cost against quality (Chen et al., 2023; Ong et al., 2025; Jung et al., 2025; Jia et al., 2026). Selective prediction adds a reject option and an explicit coverage–risk trade-off (Geifman & El-Yaniv, 2017), while LLM uncertainty signals guide answering, escalation, and abstention (Kuhn et al., 2023; Farquhar et al., 2024; Wang et al., 2025b). These signals can rank or route inputs, but neither ranking quality nor aggregate utility certifies the error rate of the answers that are ultimately returned. Reliability must therefore be defined on the joint system-level selection event.

Distribution-free calibration turns confidence scores into finite-sample guarantees: conformal prediction, probably approximately correct (PAC) confidence sets, and risk-controlling prediction sets guarantee coverage or bounded error under minimal assumptions (Angelopoulos & Bates, 2021; Bates et al., 2021). Conformal Risk Control (CRC) and Learn-then-Test extend this perspective to general losses and data-dependent selection (Angelopoulos et al., 2024; 2025). Recent foundation-model methods include ConU and SConU for conformal correctness or selective uncertainty, COIN for selective answering with risk control, and LEC for selected answers in multi-stage systems (Wang et al., 2024; 2025c; 2026b;a). These methods replace informal uncertainty ranking with guarantees for retained predictions, but they target different risk events.

This setting creates two difficulties. First, SCER need not be monotone in either threshold. Changing the Primary threshold can replace a Guardian answer with a Primary answer, whereas changing the Guardian threshold can replace either answer with abstention. Either change may therefore raise or lower the error rate among returned answers. LEC and BalanceRAG address related joint-calibration problems, while work on non-monotonic CRC shows that a safe search direction cannot be assumed (Wang et al., 2026a; Jia et al., 2026; Angelopoulos, 2026). Second, reusing certification

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.

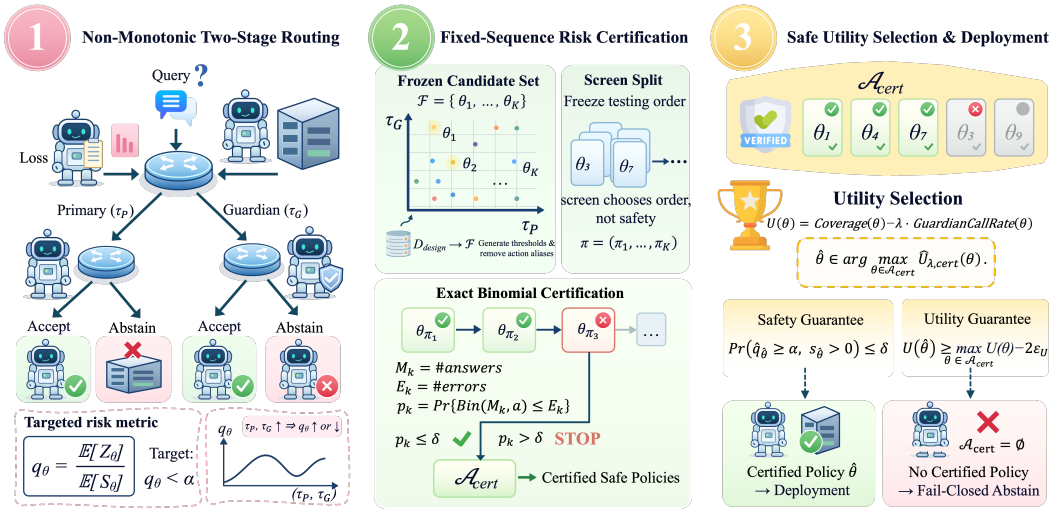


Figure 1: Overview of FS-Cert-U for non-monotonic two-stage routing. Left: Primary–Guardian–Abstain routing is controlled by two uncertainty thresholds, and system-level SCER need not vary monotonically with either threshold. Middle: design freezes a finite candidate family, screening fixes its order, and independent certification applies exact fixed-sequence testing to obtain A_{cert}. Right: utility is optimized only within A_{cert}; if the set is empty, the system fails closed.

evidence while searching over threshold pairs can invalidate fixed-policy guarantees. This connects the problem to Learn-then-Test, conformal p-value selection, and ordered multiple testing (Angelopoulos et al., 2025; Jin & Candès, 2023; Qiu et al., 2015). Together, these issues motivate certifying the candidate set before optimizing utility.

FS-Cert-U has four stages: design freezes a finite policy family; screening fixes its order; independent certification applies exact candidate-wise tests with fixed-sequence stopping; and utility optimization remains inside the certified set. If certification is empty, the system fails closed. The inferential object is the exposed set itself, so downstream utility selection does not spend an additional risk budget. Figure 1 summarizes this routing, certification, and restricted-optimization pipeline. Broader connections appear in Appendix F:

- We formalize Primary–Guardian–Abstain SCER, show why threshold monotonicity can fail, and introduce FS-Cert-U to certify a finite policy family before utility optimization.
- We derive exact candidate-wise tests and prove family-wise safety for the certified set without requiring independence across candidate tests. We also give a high-probability utility-regret bound relative to the best policy in that random set.
- Across 33 QA routing configurations and multiple risk targets, FS-Cert-U retains more utility than matched Bonferroni/Holm certification. Experiments examine non-monotonic surfaces, routing allocation, Guardian cost, and the finite-sample events covered by the theory.

2 Method

2.1 Problem formulation and routing risk

Let $W = (X, Y)$ be a target-population certification unit, including generation or judging randomness unless fixed deterministically. Fixed Primary and Guardian models produce answers, error indicators $E_P, E_G \in \{0, 1\}$ under a correctness rule fixed before calibration, and uncertainty scores $u_P(X), u_G(X)$ (smaller is more confident). For $\theta = (\tau_P, \tau_G)$, define

$$a_\theta(X) = \begin{cases} P, & u_P(X) \le \tau_P, \\ G, & u_P(X) > \tau_P \text{ and } u_G(X) \le \tau_G, \\ A, & u_P(X) > \tau_P \text{ and } u_G(X) > \tau_G, \end{cases} \tag{1}$$

where A denotes abstention. The answer and accepted-error indicators are

$$S_\theta = \mathbf{1}\{a_\theta \in \{P, G\}\}, \qquad Z_\theta = \mathbf{1}\{a_\theta = P\}E_P + \mathbf{1}\{a_\theta = G\}E_G. \tag{2}$$

Thus $Z_\theta \le S_\theta$. Let $s_\theta = \mathbb{E}[S_\theta]$ denote coverage. For $s_\theta > 0$,

$$q_\theta := \mathrm{SCER}(\theta) = \mathbb{P}\big(\text{error} \mid S_\theta = 1\big) = \frac{\mathbb{E}[Z_\theta]}{\mathbb{E}[S_\theta]}. \tag{3}$$

A policy is α-*safe* for $\alpha \in (0,1)$ precisely when

$$s_\theta > 0 \qquad \text{and} \qquad q_\theta < \alpha. \tag{4}$$

SCER is undefined at zero coverage; all-abstain is a fail-closed fallback, not a successful zero-risk policy. A certified policy must therefore return answers with positive population probability; abstention alone does not establish safety for returned answers in the routed system.

Non-monotonicity. A Primary-to-Guardian switch can correct or introduce an error, while a Guardian-to-abstain switch can remove either a correct or an incorrect answer. Relaxing the Primary threshold therefore does not simply add answered examples: it can replace a Guardian prediction with a Primary prediction on a query that was already answered. Changing the Guardian threshold also changes both which escalated queries are answered and the composition of the accepted set. Coverage may consequently increase while SCER moves in either direction, unlike ordinary one-threshold selective prediction where relaxation often yields a nested set (Wang et al., 2026a). Neither threshold is universally safe; Appendix A.1 gives the shifted-loss equivalence.

Utility. Let $C_\theta(W) = \mathbf{1}\{u_P(X) > \tau_P\}$ count Guardian calls, including calls followed by abstention. For fixed $\lambda \ge 0$, define $v_\theta^\lambda = S_\theta - \lambda C_\theta \in [-\lambda, 1]$ and

$$U_\lambda(\theta) = \mathbb{E}[v_\theta^\lambda] = \mathrm{Coverage}(\theta) - \lambda\,\mathrm{GuardianCallRate}(\theta). \tag{5}$$

Utility is optimized only after certification and therefore cannot override the safety constraint.

2.2 FS-CERT-U CONSTRUCTION

Step 1: freeze family and order. Let D_{design}, D_{screen}, D_{cert}, and D_{test} denote disjoint design, screening, certification, and test samples. Using D_{design}, construct a finite family $\mathcal{F} = \{\theta_1, \dots, \theta_K\}$; then use D_{screen} to fix a permutation $\pi = (\pi_1, \dots, \pi_K)$. Both objects are frozen before D_{cert} is inspected. Write $\mathcal{H}_{\mathrm{pre}} = \sigma(D_{\mathrm{design}}, D_{\mathrm{screen}})$, $s_k = s_{\theta_k}$, and $q_k = q_{\theta_k}$ when $s_k > 0$. Candidate IDs and deterministic tie-breaks are frozen at the same time. The exact experimental construction is in Appendix E.1; D_{test} remains held out for evaluation.

Why separate construction and ordering? The design sample determines which policies may ever be certified, whereas the screening sample determines which of those policies receive early testing priority. This distinction matters because ordering affects power but not candidate-wise validity, provided both the family and order are frozen before certification. The construction thus permits data-dependent prioritization while preventing certification outcomes from changing the tested hypothesis family.

Step 2: candidate-wise exact certification. For $m = |D_{\mathrm{cert}}|$ i.i.d. units $W_1, \dots, W_m$, define

$$M_k = \sum_{i=1}^{m} S_{\theta_k}(W_i), \qquad E_k = \sum_{i=1}^{m} Z_{\theta_k}(W_i). \tag{6}$$

Test the unsafe null

$$H_k^{\mathrm{unsafe}} : \quad s_k = 0 \quad \text{or} \quad q_k \ge \alpha. \tag{7}$$

using the one-sided exact binomial p-value (Clopper & Pearson, 1934), for $M_k > 0$,

$$p_k = \mathbb{P}\{B \le E_k : B \sim \mathrm{Binomial}(M_k, \alpha)\} = \sum_{e=0}^{E_k} \binom{M_k}{e} \alpha^e (1-\alpha)^{M_k - e}. \tag{8}$$

Set $p_k = 1$ when $M_k = 0$. Conditioning on M_k gives an exact binomial test for the frozen policy; Appendix A.2 proves that the resulting p-value is super-uniform under the unsafe null.

The denominator M_k is random because the routing policy determines which certification examples are answered. We therefore condition on the answered subset rather than replace SCER by an unconditional surrogate. Given $M_k = r$, the r answered examples have error count $E_k \sim \mathrm{Binomial}(r, q_k)$, so the lower-tail test is exact for the selection-conditioned error rate. The candidate-wise test consequently handles the random answered subset directly; Appendix A.2 gives the formal conditioning argument.

Step 3: fixed-sequence certification. Traverse π in order. Certify θ_{π_r} when $p_{\pi_r} \leq \delta$ and stop at the first failure. If failure occurs at position r, return

$$\mathcal{A}_{\mathrm{cert}} = \{\theta_{\pi_1}, \ldots, \theta_{\pi_{r-1}}\}. \tag{9}$$

The set is empty if the first test fails and equals $\mathcal{F}$ if none fails. Because the order is fixed independently of certification outcomes, each reached test may use level δ while the family-wise error rate (FWER) remains controlled. The price is that certification stops at the first failure.

Why fixed sequence? Bonferroni-type corrections protect an unordered family by dividing the error budget across candidates. FS-Cert-U instead uses the independently fixed screen order: each reached candidate is tested at the full level δ, and certification terminates at the first failure. This can improve power when the screen places certifiable, useful policies early, although a poor ordering can truncate the prefix. Thus ordering changes power without changing the validity argument.

Step 4: utility selection. For each certified policy, compute

$$\widehat{U}_{\lambda,\mathrm{cert}}(\theta_k) = \frac{1}{m}\sum_{i=1}^{m}\left(S_{\theta_k}(W_i) - \lambda C_{\theta_k}(W_i)\right). \tag{10}$$

and, when $\mathcal{A}_{\mathrm{cert}} \neq \varnothing$, select

$$\widehat{\theta} \in \arg\max_{\theta \in \mathcal{A}_{\mathrm{cert}}} \widehat{U}_{\lambda,\mathrm{cert}}(\theta), \tag{11}$$

with a measurable tie-break; otherwise, fail closed. Because the entire certified set is protected, every measurable selector restricted to $\mathcal{A}_{\mathrm{cert}}$ inherits the safety guarantee.

The same certification sample may also estimate utility after the certified set has been formed. This does not spend additional risk budget: Theorem 1 protects every member of $\mathcal{A}_{\mathrm{cert}}$ simultaneously, so utility selection cannot introduce an uncertified policy. Reusing D_{cert} does create a finite-sample utility-selection problem, which is handled uniformly by Theorem 2.

Separation of roles. The four stages isolate three forms of adaptivity. Design determines the candidate space, screening determines testing priority, and certification determines membership in the protected set. Only after this set is fixed does utility determine the deployed policy. The analysis can therefore separate two questions: whether any unsafe policy enters $\mathcal{A}_{\mathrm{cert}}$, and how much utility is lost when selecting within $\mathcal{A}_{\mathrm{cert}}$.

2.3 Finite-sample guarantees

Assumption 1 (Certification sampling and pre-certification freezing)**.** *Conditional on* $\mathcal{H}_{\mathrm{pre}} = \sigma(D_{\mathrm{design}}, D_{\mathrm{screen}})$*:*

(i) the certification units $W_1, \ldots, W_m$ are i.i.d. draws from the target population and are independent of $\mathcal{H}_{\mathrm{pre}}$;

(ii) the finite family $\mathcal{F}$, the order π, and candidate IDs are fixed; any final selector is measurable and returns an element of $\mathcal{A}_{\mathrm{cert}}$ when it is nonempty, failing closed otherwise;

(iii) α, δ, the stopping and correctness rules are fixed before certification.

Assumption 2 (Fixed utility for regret analysis)**.** *For the utility concentration and regret result, the utility definition and weight $\lambda \geq 0$ are fixed before D_{cert} is inspected. The safety theorem itself permits any measurable post-certification selector whose output remains in $\mathcal{A}_{\mathrm{cert}}$.*

The exact test follows from the conditional binomial law, with proofs in Appendix A.2. Appendix A.3 gives additional power and coverage-floor diagnostics.

Theorem 1 (Fixed-sequence family-wise safety)**.** *Under Assumption 1 and the p-values in Eq.* (8)*,*

$$\mathbb{P}(\exists \theta_k \in \mathcal{A}_{\mathrm{cert}} : \theta_k \textit{ is not } \alpha\textit{-safe} \mid \mathcal{H}_{\mathrm{pre}}) \leq \delta. \tag{12}$$

The result holds without any independence assumption across candidate p-values.

The proof tracks the first unsafe candidate in the frozen sequence. Any final selector restricted to $\mathcal{A}_{\mathrm{cert}}$ therefore inherits the same population-risk guarantee (Appendix A.6).

Theorem 1 concerns admissibility: every policy exposed to downstream optimization is safe except with family-wise probability δ. Theorem 2 concerns selection quality: conditional on obtaining a certified set, empirical utility maximization approximately matches the best population utility available within that set.

Because $v_\theta^\lambda \in [-\lambda, 1]$, Hoeffding's inequality (Hoeffding, 1963) gives a uniform bound even though certification and utility selection share data.

Theorem 2 (Uniform utility concentration and certified-set regret)**.** *Under Assumptions 1–2, let* $K = |\mathcal{F}|$ *and choose* $\eta \in (0, 1)$*. Define*

$$\varepsilon_U = (1 + \lambda)\sqrt{\frac{\log(2K/\eta)}{2m}}. \tag{13}$$

Then

$$\mathbb{P}\left(\max_{1 \leq k \leq K} \left| \widehat{U}_{\lambda,\mathrm{cert}}(\theta_k) - U_\lambda(\theta_k) \right| > \varepsilon_U \;\middle|\; \mathcal{H}_{\mathrm{pre}} \right) \leq \eta. \tag{14}$$

On the complementary event, whenever $\mathcal{A}_{\mathrm{cert}} \neq \varnothing$*,*

$$U_\lambda(\widehat{\theta}) \geq \max_{\theta \in \mathcal{A}_{\mathrm{cert}}} U_\lambda(\theta) - 2\varepsilon_U. \tag{15}$$

Scope of the utility guarantee. The regret comparison is deliberately local to the certified set. FS-Cert-U does not claim that the finite family contains the globally optimal continuous-threshold router. Candidate construction controls approximation quality, whereas Theorem 2 controls only the additional loss from finite-sample utility selection among certified candidates.

3 Experiments

Unless stated otherwise, all comparisons use 500 matched question-level splits and the same frozen candidate family. Pair-level figures use five prespecified pairs spanning model families and capacity gaps; the full list, selection rationale, and additional allocation pair are given in Appendix E.1.

3.1 Setup

We evaluate CommonsenseQA, TriviaQA, and ScienceQA on 11 directed pairs each. Unless otherwise noted, we use 10/20/35/35% design/screen/certification/test splits, $\delta = 0.05$, $\lambda = 0.25$, and $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. FS-Cert-U, Bonferroni, Holm, and Naive use the same family and matched splits. Detailed protocols, model-pair definitions, single-model references, the MM-Vet stress test, and the contextual LEC comparison are reported in Appendix E.1.

3.2 Real Non-Monotonic Routing Risk

Question. Do threshold reversals appear in actual routing surfaces? Figure 2 shows two explicit TriviaQA witnesses and one CommonsenseQA comparison from the archived held-out surfaces. Each panel plots SCER over the Primary–Guardian threshold grid and overlays the SCER $= 0.10$ contour. Along the stronger TriviaQA Guardian-direction path, SCER follows $0.383 \to 0.176 \to 0.375$; along the weaker Primary-direction path, it follows $0.0323 \to 0.0275 \to 0.0350$. Additional held-out surfaces across datasets and pairs appear in Appendix E.9.

These finite-pool displays do not establish global population non-monotonicity. They do show that the routing geometry can contain threshold reversals, so the architecture offers no universally safe monotone search direction. We therefore next certify a frozen routing family directly.

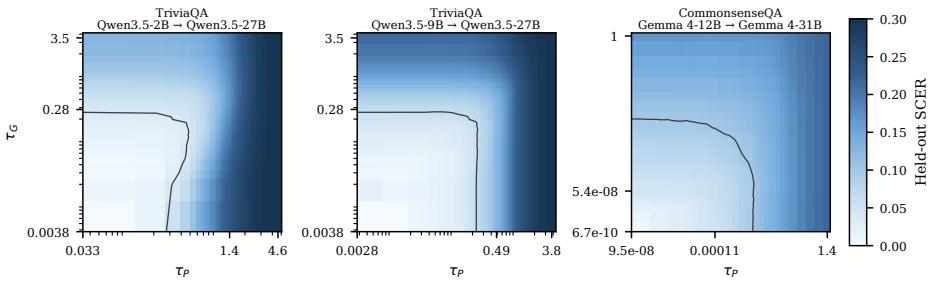


Figure 2: Three archived held-out SCER surfaces. The two TriviaQA panels show the reported strong Guardian-direction and weaker Primary-direction finite-pool witnesses; the CommonsenseQA panel is comparative. The contour marks SCER $= 0.10$; none establishes global population non-monotonicity. Additional surfaces appear in Appendix E.9.

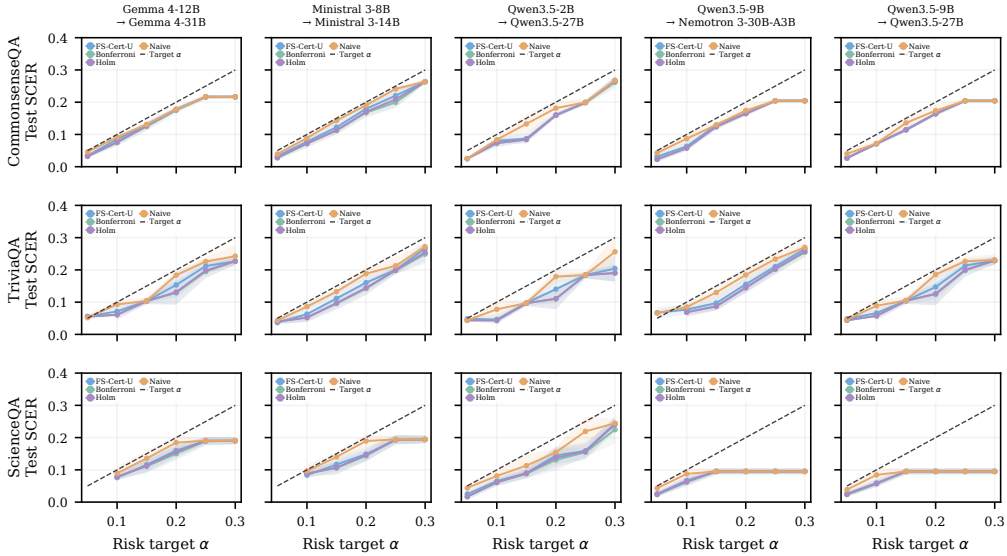


Figure 3: System-level SCER across three datasets and five representative routing pairs. Curves show 500-split means ± SD for matched methods using the same frozen family.

3.3 System-Level Risk Control

Question. Does FS-Cert-U operate in the intended system-risk regime across pairs, datasets, and targets? Figure 3 reports test SCER for five fixed representative pairs on each QA dataset.

FS-Cert-U operates empirically closer to the target than the more conservative Bonferroni and Holm baselines while retaining the same frozen candidate family. These held-out SCER curves describe operating behavior; the family-wise certification event controlled by Theorem 1 is evaluated separately in Section 3.7. Naive can retain more useful behavior but provides no family-wise certificate. The pattern appears across datasets and pair types rather than being a dataset-average artifact. Full 11-pair curves and exceedance diagnostics remain in Appendix E.3.

3.4 Certification Power and Its Consistency

Question. Is the certification-power gain confined to a particular risk target or a small subset of routing pairs? Within the same frozen family, we measure certification power by the useful policy space retained after certification: coverage and utility, with SCER tracking the resulting operating risk. Table 1 summarizes the four matched methods on these four core system metrics at $\alpha = 0.10$, $\lambda = 0.25$, averaged over all 11 pairs per dataset.

Single-model references, nonempty rates, and accepted-correct counts are reported in Appendix E.2.

At the primary operating point, FS-Cert-U has higher utility than Bonferroni and Holm on all three datasets; Naive can be higher, but it lacks the certificate. Figure 4 examines both dimensions. The

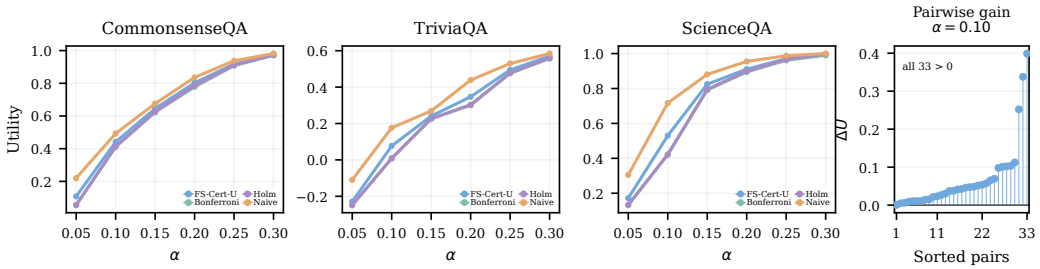


Figure 4: Certification power across risk targets and routing pairs. The first three panels show utility across all six risk targets at $\lambda = 0.25$, averaged equally over 11 pairs per dataset and 500 matched splits; the compact fourth panel shows the sorted mean utility difference $\Delta U = U_{\text{FS-Cert-U}} - U_{\text{Bonferroni}}$ for all 33 pairs at $\alpha = 0.10$, with the zero line in black. Complete pair labels and percentile intervals appear in Appendix Fig. 24.

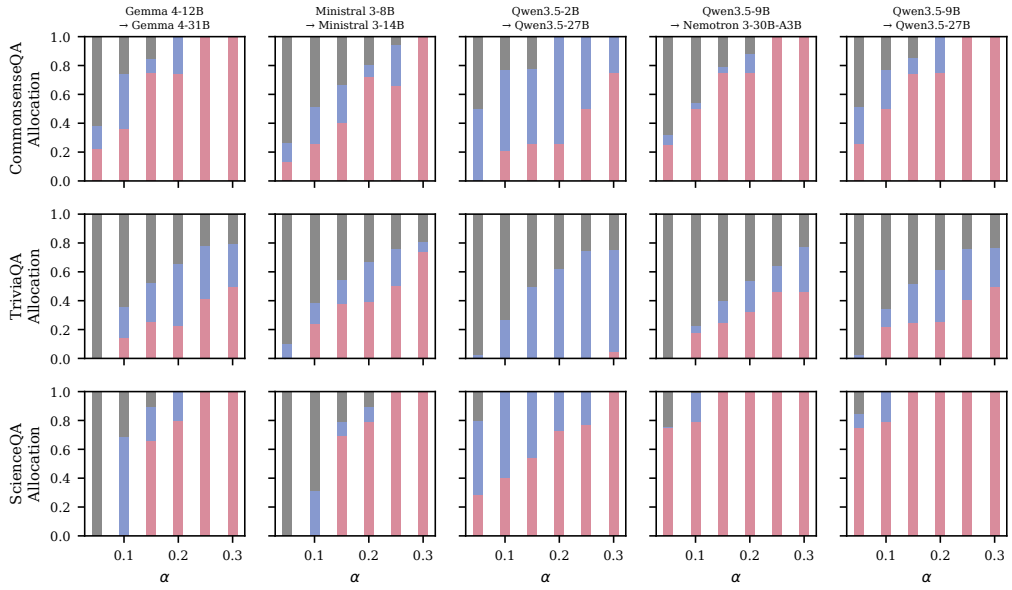


Figure 5: FS-Cert-U allocation across three datasets and five representative pairs. Stacked fractions show Primary answers (pink), Guardian answers (blue), and abstention (gray) over risk targets.

first three panels show that the dataset-level utility advantage over Bonferroni persists across all six risk targets, while the fourth shows positive paired mean differences for all 33 routing configurations at $\alpha = 0.10$. Complete pair labels and empirical percentile intervals are reported in Appendix Fig. 24. These results support greater certification power within the matched frozen-family comparison, rather than universal routing dominance.

3.5 Routing Allocation and Correct-Answer Recovery

Question. What policies does utility optimization actually select? Figure 5 displays FS-Cert-U allocations across five fixed pairs and three datasets, separating Primary answers, Guardian answers, and abstentions over α.

The method does not force both stages to be used. For Qwen3.5-2B→Qwen3.5-9B on TriviaQA at $\alpha = 0.10$, the Primary answered fraction is zero and accepted answers

Table 1: Core matched comparison at $\alpha = 0.10$, $\lambda = 0.25$. Means are equal-weight averages over 11 pairs and 500 matched splits per dataset.

Dataset	Method	SCER (%)	Coverage (%)	Guardian calls (%)	Utility
CommonsenseQA	Naive	8.66	64.84	62.14	0.493
	Bonferroni	6.73	59.32	72.89	0.411
	Holm	6.77	59.51	72.55	0.414
	FS-Cert-U	7.41	61.41	69.11	0.441
TriviaQA	Naive	8.32	39.09	85.82	0.176
	Bonferroni	6.07	23.24	89.66	0.008
	Holm	6.09	23.31	89.61	0.009
	FS-Cert-U	6.86	29.40	86.88	0.077
ScienceQA	Naive	9.03	84.95	53.30	0.716
	Bonferroni	6.08	59.27	68.86	0.421
	Holm	6.14	59.39	68.63	0.422
	FS-Cert-U	6.80	69.75	67.11	0.530

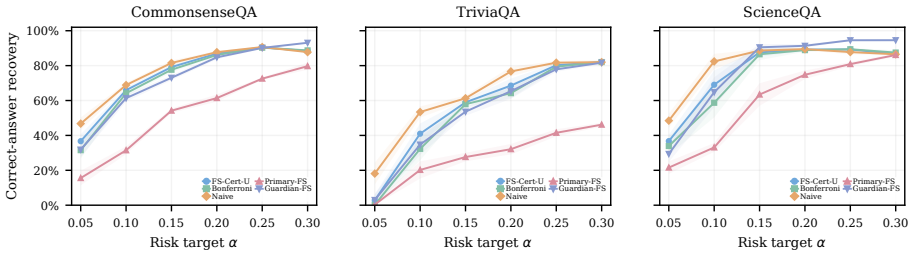


Figure 6: Correct-answer recovery across all 11 pairs per dataset. Recovery is a retention measure: accepted correct answers divided by examples for which at least one cached model is correct.

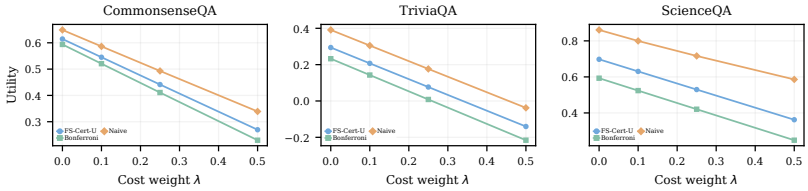


Figure 7: Utility versus Guardian cost weight λ at $\alpha = 0.10$, averaged over 11 pairs per dataset and 500 matched splits. FS-Cert-U retains a positive utility difference over Bonferroni at all four prespecified costs; the TriviaQA value becomes negative at $\lambda = 0.50$.

are assigned to the Guardian branch.
Other pairs share queries between the two stages when that improves certified utility. Figure 6 reports correct-answer recovery across all 11 pairs per dataset, defined as accepted correct answers divided by test examples for which at least one cached model is correct. Higher recovery measures retention, whereas SCER remains the risk metric.

3.6 Cost Robustness

Question. Does the certification-power advantage persist as Guardian cost changes? Figure 7 varies $\lambda \in \{0, 0.10, 0.25, 0.50\}$ at $\alpha = 0.10$ and reports utility on all three datasets. Appendix Fig. 30 gives the corresponding Guardian-call-rate/coverage operating points.

The FS-Cert-U minus Bonferroni utility difference remains positive at all four prespecified costs. At $\lambda = 0.50$, TriviaQA FS-Cert-U utility is negative (-0.1404): when escalation is sufficiently expensive, no certification method can manufacture a favorable operating point. The appendix operating points are discrete; they do not form a continuous Pareto frontier or compare methods at identical call budgets. Because benchmark behavior depends on escalation cost, we next test the theorem events directly.

3.7 Theorem-Aligned Validation

Question. Do the random events in Theorems 1 and 2 behave as predicted? Figure 8 combines the controlled-population replay. Panel (a) estimates false certification for FS-Cert-U, Bonferroni, and

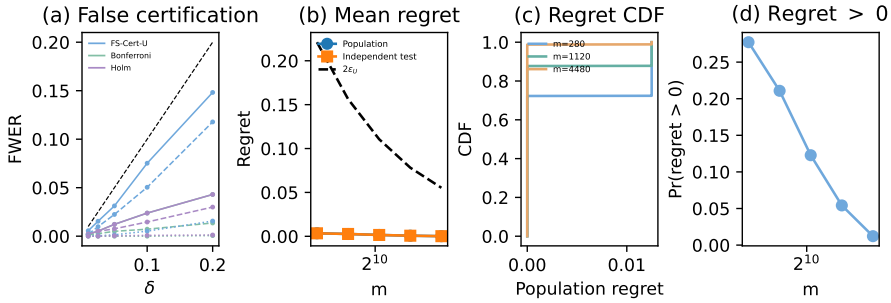


Figure 8: Controlled-population validation. (a) False certification versus δ for FS-Cert-U, Bonferroni, and Holm. (b) Mean population and independent-test regret with $2\varepsilon_U$. (c) Regret CDFs at three certification sizes. (d) Probability of positive regret versus certification size.

Holm; it targets the event that an unsafe policy enters the certified set, rather than finite-test SCER. Panels (b)–(d) use a separate near-tie population with $m \in \{280, 560, 1120, 2240, 4480\}$.

Mean population regret decreases from 0.003465 at $m = 280$ to 0.0001525 at $m = 4480$, while the probability of positive regret falls from 27.72% to 1.22%. The analytical $2\varepsilon_U$ bound remains conservative. Thus the controlled experiment supports the theorem's finite-sample trend rather than its numerical tightness; population utilities and replay construction are specified in Appendix E.9.

3.8 What Determines Certification Power?

Question. Which design choices affect power without changing validity? Figure 9 summarizes archived sensitivity studies across the main configurations. Screen ordering changes the certified prefix, certification evidence changes utility and the nonempty rate, and candidate density changes attainable utility while SCER remains on the same scale.

These factors determine power without changing the fixed-sequence validity argument. Full ordering, certification-fraction, candidate-density, and split-level results are in Appendix E.4; MM-Vet v2 provides an additional stress test with a small certification set in Appendix E.7.

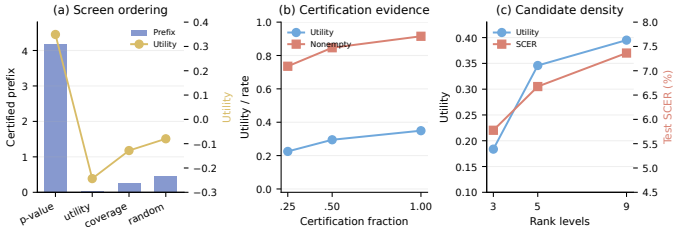


Figure 9: Archived power mechanisms across the main configurations. Ordering, certification evidence, and candidate density affect power and attainable utility; validity is supplied by the frozen-family certification construction.

4 Conclusion

FS-Cert-U certifies a finite policy set before optimizing utility. By freezing the family and its order and applying exact fixed-sequence tests, it obtains set-level population safety without requiring monotone routing risk. In matched experiments, it retains more utility than generic multiplicity corrections, with the advantage over Bonferroni persisting across Guardian costs. Controlled replays separately validate false-certification control and utility regret within the certified set. These results support the construction within the matched comparison class, rather than universal optimality over all routing rules.

Together, the held-out surfaces explain why monotone threshold search is insufficient, the matched comparisons quantify certification power, and the controlled replay isolates the safety and utility guarantees. Because the family, order, certification event, and utility choice remain explicit, the method is auditable and clarifies whether future gains require a richer family, a better order, or a new validity argument.

4.1 Limitations

FS-Cert-U is limited to a finite frozen family, so candidate richness, screening order, and certification sample size affect power. Its exact test requires i.i.d. certification units; clustered data need valid group-level units or another candidate-wise test. The correctness rule, judge, and generation randomness must match deployment, while distribution shift, subgroup guarantees, and repeated deployment require extensions. Richer external searches may yield higher utility under different guarantees. Finally, the safety statement concerns population risk over certification randomness and does not require every finite test split to fall below the target.

AI USE STATEMENT

Generative AI assisted drafting, editing, and LaTeX formatting. The authors take responsibility for the final content and verification of all claims, proofs, citations, and experimental results.

REPRODUCIBILITY STATEMENT

Appendix E specifies the protocols and supplementary artifacts. The final release will include the data and code needed to reproduce candidate construction, certification, selection, and evaluation.

REFERENCES

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. *arXiv preprint arXiv:2602.20151*, 2026. URL https://arxiv.org/abs/2602.20151.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://research.google/pubs/conformal-risk-control/.

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.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009. doi: 10.1002/sim.3495.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023. URL https://arxiv.org/abs/2305.05176.

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.

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.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979.

Anqi Hu, Zhiyuan Wang, Zijun Jia, and Bo Fu. MiRD: Reliable set-valued prediction for open-ended question answering via miscoverage risk decomposition. *arXiv preprint arXiv:2605.27091*, 2026. URL https://arxiv.org/abs/2605.27091.

Zijun Jia, Yuanchang Ye, Sen Jia, Yiyao Qian, Haoning Wang, Baojie Chen, Diyin Tang, Jinsong Yu, and Zhiyuan Wang. BalanceRAG: Joint risk calibration for cascaded retrieval-augmented generation. *arXiv preprint arXiv:2605.20084*, 2026. URL https://arxiv.org/abs/2605.20084.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2210.01408.

Mandar Joshi, Eunsol Choi, Daniel S. 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*, pp. 1601–1611, 2017.

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://openreview.net/forum?id=UHPnqSTBPO.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026. URL https://arxiv.org/abs/2601.08584.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

NVIDIA. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026. URL https://arxiv.org/abs/2604.24954.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxVI04YvB.

Zhiying Qiu, Wenge Guo, and Gavin Lynch. On generalized fixed sequence procedures for controlling the FWER. *Statistics in Medicine*, 34(30):3968–3983, 2015. doi: 10.1002/sim.6603.

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.

Qwen Team. Qwen3.8-27B: Model card. Hugging Face model repository, 2026a. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pp. 4149–4158, 2019.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. *arXiv preprint arXiv:2510.10193*, 2025a. URL https://arxiv.org/abs/2510.10193.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.404.

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 139:109553, 2025b. doi: 10.1016/j.engappai.2024.109553.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19052–19075. Association for Computational Linguistics, 2025c. doi: 10.18653/v1/2025.acl-long.934.

Zhiyuan Wang, Aniri, Tianlong Chen, Yue Zhang, Heng Tao Shen, Xiaoshuang Shi, and Kaidi Xu. LEC: Linear expectation constraints for selection-conditioned risk control in selective prediction and routing systems. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a. URL https://arxiv.org/abs/2512.01556.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. COIN: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33764–33772, 2026b. doi: 10.1609/aaai.v40i40.40667.

W. Yu, Z. Yang, L. Ren, L. Li, J. Wang, K. Lin, C.-C. Lin, Z. Liu, L. Wang, and X. Wang. MM-Vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.

A Detailed Proofs

A.1 SCER and the shifted linear loss

Following the selection/error linearization, define

$$\ell_{\alpha,\theta}(W) = Z_\theta(W) - \alpha S_\theta(W) + \alpha. \tag{A.1}$$

Proposition 1 (Equivalence of shifted loss and SCER)**.** *For any fixed policy θ with $s_\theta = \mathbb{E}[S_\theta] > 0$,*

$$\mathbb{E}[\ell_{\alpha,\theta}] \le \alpha \quad \Longleftrightarrow \quad q_\theta \le \alpha. \tag{A.2}$$

Proof. By Eq. (A.1),

$$\mathbb{E}[\ell_{\alpha,\theta}] = \mathbb{E}[Z_\theta] - \alpha\mathbb{E}[S_\theta] + \alpha. \tag{A.3}$$

Therefore $\mathbb{E}[\ell_{\alpha,\theta}] \le \alpha$ if and only if

$$\mathbb{E}[Z_\theta] \le \alpha\mathbb{E}[S_\theta]. \tag{A.4}$$

Because $s_\theta = \mathbb{E}[S_\theta] > 0$, division by s_θ gives

$$\frac{\mathbb{E}[Z_\theta]}{\mathbb{E}[S_\theta]} \le \alpha, \tag{A.5}$$

which is exactly $q_\theta \le \alpha$ by Eq. (3). The reverse direction is the same algebra in reverse. □

Why this does not create monotonicity. The transformation is algebraic only. It turns a conditional ratio target into a bounded expected-loss inequality, but it does not impose an ordering on θ. The examples in Section 2.1 show that the sample-wise loss can move in both directions under the same threshold change.

A.2 Conditional binomial law and exact p-values

Lemma 1 (Conditional binomial law)**.** *Fix a candidate θ_k conditional on $\mathcal{H}_{\mathrm{pre}}$. If $s_k > 0$, then*

$$E_k \mid (M_k = r, \mathcal{H}_{\mathrm{pre}}) \sim \mathrm{Binomial}(r, q_k), \qquad r = 0, 1, \ldots, m. \tag{A.6}$$

Lemma 2 (Super-uniform exact p-value)**.** *Under Assumption 1, for any frozen candidate θ_k satisfying the unsafe null in Eq. (7),*

$$\mathbb{P}(p_k \le t \mid \mathcal{H}_{\mathrm{pre}}) \le t, \qquad 0 \le t \le 1. \tag{A.7}$$

We now prove Lemmas 1 and 2. Fix $\mathcal{H}_{\mathrm{pre}}$ throughout, so θ_k is deterministic and the certification units are i.i.d.

Proof of Lemma 1. For a fixed candidate, each certification unit produces the pair (S_i, Z_i) with support

$$(0,0), \qquad (1,0), \qquad (1,1). \tag{A.8}$$

Write

$$s_k = \mathbb{P}(S_i = 1), \qquad q_k = \mathbb{P}(Z_i = 1 \mid S_i = 1). \tag{A.9}$$

Then the three cell probabilities are

$$1 - s_k, \qquad s_k(1 - q_k), \qquad s_k q_k. \tag{A.10}$$

Across m i.i.d. certification units, the counts

$$N_0 = \#\{S_i = 0\}, \quad N_c = \#\{S_i = 1, Z_i = 0\}, \quad N_e = \#\{S_i = 1, Z_i = 1\} \tag{A.11}$$

are multinomial with the probabilities above. Moreover

$$M_k = N_c + N_e, \qquad E_k = N_e. \tag{A.12}$$

Conditioning a multinomial on $M_k = r$ leaves the r selected units split between correct and incorrect with normalized probabilities $1 - q_k$ and q_k. Hence

$$E_k \mid (M_k = r, \mathcal{H}_{\text{pre}}) \sim \text{Binomial}(r, q_k). \tag{A.13}$$

This includes $r = 0$ as a degenerate binomial. □

Proof of Lemma 2. First consider $s_k = 0$. Then $M_k = 0$ almost surely, so by definition $p_k = 1$ and

$$\mathbb{P}(p_k \le t \mid \mathcal{H}_{\text{pre}}) = 0 \le t \tag{A.14}$$

for every $t < 1$; the claim is trivial at $t = 1$.

Now suppose $s_k > 0$ and $q_k \ge \alpha$. Conditional on $M_k = r > 0$, Lemma 1 gives $E_k \sim \text{Binomial}(r, q_k)$. Define the reference lower-tail CDF

$$F_{r,\alpha}(e) = \mathbb{P}\{\text{Binomial}(r, \alpha) \le e\}. \tag{A.15}$$

The p-value is $p_k = F_{r,\alpha}(E_k)$. For any $t \in [0, 1]$, let

$$c_t = \max\{e \in \{0, \dots, r\} : F_{r,\alpha}(e) \le t\}, \tag{A.16}$$

with the event empty if no such e exists. Then

$$\begin{aligned} \mathbb{P}(p_k \le t \mid M_k = r, \mathcal{H}_{\text{pre}}) &= \mathbb{P}_{q_k}(E_k \le c_t) && \text{(A.17)} \\ &\le \mathbb{P}_{\alpha}(E_k \le c_t) && \text{(A.18)} \\ &= F_{r,\alpha}(c_t) && \text{(A.19)} \\ &\le t. && \text{(A.20)} \end{aligned}$$

The inequality uses stochastic monotonicity of a binomial count in its success probability: for a fixed lower-tail threshold, the probability of a small count is non-increasing as q increases, and $q_k \ge \alpha$.

For $M_k = 0$, $p_k = 1$ and the same bound holds. Averaging the conditional inequality over the random M_k proves

$$\mathbb{P}(p_k \le t \mid \mathcal{H}_{\text{pre}}) \le t. \tag{A.21}$$

□

A.3 Certifiability and coverage-floor calculations

Proposition 2 (Certifiability with zero observed errors)**.** *If a candidate answers $M > 0$ certification units and makes no observed errors ($E = 0$), then its exact p-value is*

$$p = (1 - \alpha)^M. \tag{A.22}$$

Consequently, zero observed errors can certify the candidate at level δ only if

$$M \ge M_{\min}(\alpha, \delta) := \left\lceil \frac{\log \delta}{\log(1 - \alpha)} \right\rceil. \tag{A.23}$$

This is a necessary observed-answer count, not a guarantee that a candidate will answer that many units.

The calculation makes the power–coverage trade-off explicit: small target risks and small error budgets require many answered certification examples even in the most favorable zero-error outcome.

Optional minimum-coverage certification. The default safety definition requires only positive coverage because the main theorem focuses on selection-conditioned error. In applications where near-all-abstention policies are unacceptable, choose a deployment floor $s_{\min} \in (0, 1)$ and certify the stronger property $s_k \ge s_{\min}$ together with $q_k < \alpha$. For a candidate, use the upper-tail coverage p-value

$$p_k^{\text{cov}} = \mathbb{P}\{B \ge M_k : B \sim \text{Binomial}(m, s_{\min})\}, \tag{A.24}$$

which tests the unsafe coverage null $s_k < s_{\min}$. Define

$$p_k^{\text{joint}} = \max\{p_k, p_k^{\text{cov}}\}. \tag{A.25}$$

Because the joint unsafe null is the union of the risk and coverage nulls, p_k^{joint} is super-uniform whenever either component null is true. The fixed-sequence procedure is unchanged after replacing p_k by p_k^{joint}. The resulting set-level theorem certifies $s_k \geq s_{\min}$ and $q_k < \alpha$ simultaneously for every accepted candidate. This extension is optional; all main-text results with the original rule use the fail-closed positive-coverage definition.

For Proposition 2, setting $E = 0$ in Eq. (8) gives $p = (1-\alpha)^M$; solving $p \leq \delta$ yields Eq. (A.23) because both logarithms are negative.

For the optional coverage floor, $M_k \sim \text{Binomial}(m, s_k)$. Under $s_k < s_{\min}$, the upper-tail probability in Eq. (A.24) is super-uniform by binomial stochastic monotonicity, with the least favorable boundary at $s_k = s_{\min}$. Under the union unsafe null, at least one of p_k or p_k^{cov} is super-uniform; therefore $\max\{p_k, p_k^{\text{cov}}\}$ is super-uniform because the event that its maximum is at most t is contained in the event that the valid component is at most t.

A.4 Power bound

Power diagnostic under a safety margin. The FWER theorem controls false certification; a complementary margin calculation describes when a genuinely safe policy is likely to pass. Suppose a fixed candidate satisfies $s_k \geq s_0 > 0$ and $q_k \leq \alpha - \gamma$ for some $\gamma \in (0, \alpha)$. Define

$$r_\delta = \left\lceil \frac{2\log(1/\delta)}{\gamma^2} \right\rceil. \tag{A.26}$$

If $ms_0/2 \geq r_\delta$, then

$$\mathbb{P}(p_k > \delta \mid \mathcal{H}_{\text{pre}}) \leq \exp\Big(-\frac{ms_0}{8}\Big) + \delta. \tag{A.27}$$

Thus a positive coverage floor and a nonzero risk margin jointly give a usable lower bound on the probability that a safe candidate is certified. For a safe prefix, a union bound over its candidates gives a corresponding lower bound on reaching the end of that prefix. The bound is intentionally conservative and is a power diagnostic rather than an additional assumption in the safety theorem.

Let $M = M_k$. Since $M \sim \text{Binomial}(m, s_k)$ and $s_k \geq s_0$, the multiplicative Chernoff bound gives

$$\mathbb{P}(M < ms_0/2) \leq \exp(-ms_0/8). \tag{A.28}$$

On the event $M \geq ms_0/2 \geq r_\delta$, conditional binomial Hoeffding gives, under $q_k \leq \alpha - \gamma$,

$$\mathbb{P}\left(\frac{E_k}{M} > \alpha - \frac{\gamma}{2} \,\middle|\, M\right) \leq \exp(-M\gamma^2/2) \leq \delta. \tag{A.29}$$

If $E_k/M \leq \alpha - \gamma/2$, the exact lower-tail p-value is at most the corresponding binomial tail bound $\exp(-M\gamma^2/2)$ and hence is at most δ. Combining the two failure events proves Eq. (A.27).

A.5 Fixed-sequence family-wise safety

Proof of Theorem 1. Condition on $\mathcal{H}_{\text{pre}}$. The candidate family and ordering are fixed. If every candidate is α-safe, the event in Eq. (12) is impossible. Otherwise define

$$J = \min\{r \in \{1, \ldots, K\} : \theta_{\pi_r} \text{ is not } \alpha\text{-safe}\}. \tag{A.30}$$

Because safety is a population property of a frozen candidate, J is fixed conditional on $\mathcal{H}_{\text{pre}}$.

If $\mathcal{A}_{\text{cert}}$ contains any unsafe candidate, then the first unsafe candidate θ_{π_J} must itself have entered $\mathcal{A}_{\text{cert}}$. Under the fixed-sequence rule this requires, in particular,

$$p_{\pi_J} \leq \delta. \tag{A.31}$$

Therefore

$$\mathbb{P}(\exists \theta_k \in \mathcal{A}_{\text{cert}} : \theta_k \text{ is not } \alpha\text{-safe} \mid \mathcal{H}_{\text{pre}}) \tag{A.32}$$

$$\leq \mathbb{P}(p_{\pi_J} \leq \delta \mid \mathcal{H}_{\text{pre}}) \leq \delta, \tag{A.33}$$

where the last inequality is Lemma 2. No joint distributional assumption on the vector $(p_1, \ldots, p_K)$ was used. □

Unconditional form. Taking expectation with respect to $\mathcal{H}_{\text{pre}}$ yields

$$\mathbb{P}(\exists \theta_k \in \mathcal{A}_{\text{cert}} : \theta_k \text{ is not } \alpha\text{-safe}) = \mathbb{E}[\mathbb{P}(\text{same event} \mid \mathcal{H}_{\text{pre}})] \leq \delta. \tag{A.34}$$

A.6 UTILITY SELECTION DOES NOT SPEND ADDITIONAL RISK BUDGET

Corollary 1 (Safety of the utility-selected policy)**.** *Let $\widehat{\theta}$ be the output of Eq.* (11) *when $\mathcal{A}_{\mathrm{cert}} \neq \varnothing$, and fail closed otherwise. Then*

$$\mathbb{P}\big(s_{\widehat{\theta}} > 0, q_{\widehat{\theta}} \geq \alpha \;\big|\; \mathcal{H}_{\mathrm{pre}}\big) \leq \delta. \tag{A.35}$$

Consequently the same inequality holds unconditionally after averaging over design and screen randomness.

Proof of Corollary 1. When $\mathcal{A}_{\mathrm{cert}} = \varnothing$, the procedure fails closed and does not output a positive-coverage routing policy. When $\mathcal{A}_{\mathrm{cert}} \neq \varnothing$, the selector in Eq. (11) always satisfies $\widehat{\theta} \in \mathcal{A}_{\mathrm{cert}}$. Hence the event

$$\{\widehat{\theta} \text{ is unsafe}\} \tag{A.36}$$

is a subset of

$$\{\exists \theta_k \in \mathcal{A}_{\mathrm{cert}} : \theta_k \text{ is unsafe}\}. \tag{A.37}$$

Theorem 1 therefore gives Eq. (A.35). The proof does not depend on how the utility maximizer is chosen inside $\mathcal{A}_{\mathrm{cert}}$, beyond the requirement that it cannot introduce a policy outside the certified set. □

A.7 UNIFORM UTILITY CONCENTRATION AND REGRET

Proof of Theorem 2. Condition on $\mathcal{H}_{\mathrm{pre}}$. For each fixed candidate θ_k, the certification utilities

$$v_{k,i} = S_{\theta_k}(W_i) - \lambda C_{\theta_k}(W_i) \tag{A.38}$$

are i.i.d., have mean $U_\lambda(\theta_k)$, and lie in an interval of length $1 + \lambda$. Hoeffding's inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\widehat{U}_{\lambda,\mathrm{cert}}(\theta_k) - U_\lambda(\theta_k)\Big| > \varepsilon \mid \mathcal{H}_{\mathrm{pre}}\Big) \leq 2\exp\Big(-\frac{2m\varepsilon^2}{(1+\lambda)^2}\Big). \tag{A.39}$$

A union bound over the K frozen candidates yields

$$\mathbb{P}\Big(\max_k \Big|\widehat{U}_{\lambda,\mathrm{cert}}(\theta_k) - U_\lambda(\theta_k)\Big| > \varepsilon \mid \mathcal{H}_{\mathrm{pre}}\Big) \leq 2K\exp\Big(-\frac{2m\varepsilon^2}{(1+\lambda)^2}\Big). \tag{A.40}$$

Setting the right-hand side equal to η gives Eq. (13) and proves Eq. (14).

Now work on the uniform-concentration event and suppose $\mathcal{A}_{\mathrm{cert}} \neq \varnothing$. Let

$$\theta^* \in \arg\max_{\theta \in \mathcal{A}_{\mathrm{cert}}} U_\lambda(\theta). \tag{A.41}$$

Although $\mathcal{A}_{\mathrm{cert}}$ is random and depends on the same certification data, the concentration event holds simultaneously for every candidate in the entire frozen family, so it also holds for the data-dependent members $\widehat{\theta}$ and θ^*. Therefore

$$U_\lambda(\widehat{\theta}) \geq \widehat{U}_{\lambda,\mathrm{cert}}(\widehat{\theta}) - \varepsilon_U \tag{A.42}$$

$$\geq \widehat{U}_{\lambda,\mathrm{cert}}(\theta^*) - \varepsilon_U \tag{A.43}$$

$$\geq U_\lambda(\theta^*) - 2\varepsilon_U, \tag{A.44}$$

where the middle inequality follows from the definition of $\widehat{\theta}$. This is Eq. (15). □

A.8 JOINT SAFETY–UTILITY STATEMENT

Corollary 2 (Joint safety–utility statement)**.** *Under Assumptions 1–2, with conditional probability at least $1 - \delta - \eta$, either FS-Cert-U fails closed or it returns a policy satisfying both*

$$q_{\widehat{\theta}} < \alpha \tag{A.45}$$

and

$$U_\lambda(\widehat{\theta}) \geq \max_{\theta \in \mathcal{A}_{\mathrm{cert}}} U_\lambda(\theta) - 2\varepsilon_U. \tag{A.46}$$

Proof of Corollary 2. Let $\mathcal{E}_R$ denote the event that $\mathcal{A}_{\mathrm{cert}}$ contains an unsafe policy and let $\mathcal{E}_U$ denote the failure of the uniform utility concentration event. Theorem 1 gives

$$\mathbb{P}(\mathcal{E}_R \mid \mathcal{H}_{\mathrm{pre}}) \leq \delta, \tag{A.47}$$

and Theorem 2 gives

$$\mathbb{P}(\mathcal{E}_U \mid \mathcal{H}_{\mathrm{pre}}) \leq \eta. \tag{A.48}$$

By the union bound,

$$\mathbb{P}(\mathcal{E}_R \cup \mathcal{E}_U \mid \mathcal{H}_{\mathrm{pre}}) \leq \delta + \eta. \tag{A.49}$$

Outside this union, every certified policy is safe and the utility bound holds simultaneously. Thus, if $\mathcal{A}_{\mathrm{cert}}$ is nonempty, the returned policy is safe and satisfies Eq. (15); if it is empty, the procedure fails closed. □

B Assumptions, Failure Modes, and Practical Boundaries

The formal guarantee depends on a small number of operational rules. They are part of the method, not implementation details.

What must be frozen before certification. For safety, the candidate family, candidate IDs, alias-removal rule, fixed-sequence ordering, target α, family-wise error level δ, stopping rule, and correctness rule must all be fixed before certification outcomes are inspected. The utility weight λ and utility definition must additionally be fixed before certification when invoking the utility-regret theorem; the final tie-break may be any measurable rule that returns an element of the certified set. Looking at certification errors and then adding a threshold, deleting a failed candidate, or reordering the sequence invalidates the proof unless a new multiplicity argument is supplied.

Test data must not feed back into calibration. The test split is for reporting SCER, coverage, utility, call rate, action allocation, and empirical violation frequency. It must not be used to choose the best α, λ, candidate family, screen rule, or final policy. The utility-regret theorem additionally assumes that λ was fixed before certification.

Repeated occurrences and groups. If several observations correspond to the same underlying question, user, document, or other dependent unit, they should not be scattered across splits and treated as independent Bernoulli trials. The implementation uses group-level paired splitting. More generally, the theorem applies at the level at which the certification units are i.i.d.

Zero answers. When $M_k = 0$, FS-Cert-U sets $p_k = 1$. A zero-answer candidate therefore cannot be positively certified from that split. If no candidate is certified, the system may still fail closed by abstaining, but the event is reported as an empty certification set rather than as a successful SCER of zero.

Proxy correctness labels. The theorem controls the supplied binary error label. If semantic similarity or an automated judge is used as a proxy for correctness, the resulting statement concerns that proxy unless an additional argument links it to the desired semantic notion of correctness.

Alternative dependence structures. The fixed-sequence theorem itself only needs each candidate p-value to be super-uniform under its unsafe null. Therefore FS-Cert-U is modular: if a dependence-aware or group-aware candidate-level test is available, it may replace the binomial p-value while the set-level fixed-sequence proof remains unchanged. What cannot be done is to invoke arbitrary exchangeability and silently retain the i.i.d. binomial calculation.

C Computational Notes

Stage I: freeze a finite candidate family. The design stage constructs

$$\mathcal{F} = \{\theta_1, \ldots, \theta_K\}, \qquad \theta_k = (\tau_{P,k}, \tau_{G,k}). \tag{C.1}$$

For compactness, write $s_k := s_{\theta_k}$ and $q_k := q_{\theta_k}$ for the population coverage and SCER of candidate θ_k, respectively. Thus the candidate-indexed quantities below use the population definitions in Eq. (3), with q_k defined when $s_k > 0$. Our implementation uses pre-specified Primary and Guardian score quantiles computed on D_{design}, adds infinite endpoint thresholds, and forms their Cartesian product. It removes duplicate threshold values and retains a single always-Primary policy, for which the Guardian threshold is irrelevant on every input. Appendix E.1 specifies the exact grid used in the experiments. More general design-stage action-alias removal is permissible, but equality of actions on design data alone does not establish population equivalence.

The theorem does not require this particular construction. It only requires $K < \infty$ and that $\mathcal{F}$, candidate IDs, and utility tie-breaking rules be fixed before D_{cert} is inspected. In particular, no candidate may be added, removed, or merged after observing certification errors.

Stage II: freeze the fixed-sequence order. Using only D_{screen}, assign every candidate a pre-specified priority score and produce a permutation

$$\pi = (\pi_1, \ldots, \pi_K). \tag{C.2}$$

For example, the priority can combine a screen-only provisional score $\widetilde{p}_k^{\mathrm{screen}}$, screen utility, and finally candidate ID as a deterministic tie-break. We reserve p_k for the independent certification p-value. The exact ranking rule is not part of the validity theorem. The only statistical requirement is that the order be $\mathcal{H}_{\mathrm{pre}}$-measurable and frozen before the certification data are read.

FS-Cert-U procedure.

1. **Input:** $D_{\text{design}}, D_{\text{screen}}, D_{\text{cert}}$; target α; error budget δ; utility weight λ.
2. Build a finite candidate family $\mathcal{F}$ on D_{design}; remove duplicate thresholds and redundant always-Primary policies; freeze candidate IDs and ordering tie-breaks.
3. Use only D_{screen} to compute and freeze an ordering $\pi = (\pi_1, \ldots, \pi_K)$.
4. Initialize $\mathcal{A}_{\text{cert}} = \varnothing$. Traverse $r = 1, \ldots, K$. For $k = \pi_r$, compute (M_k, E_k) on D_{cert} and set $p_k = 1$ if $M_k = 0$, otherwise use Eq. (8).
5. Stop at the first $p_k > \delta$; otherwise add θ_k to $\mathcal{A}_{\text{cert}}$ and continue.
6. If $\mathcal{A}_{\text{cert}} = \varnothing$, return fail-closed abstention. Otherwise return the maximizer of Eq. (11) over $\mathcal{A}_{\text{cert}}$.

Figure 10: The FS-Cert-U pipeline. If the optional coverage floor in Appendix A.3 is used, replace p_k in step 4–5 by the joint p-value p_k^{joint}.

This split has a clear purpose. Fixed-sequence testing stops at the first failed certification, so promising policies should appear early. Screen data can therefore improve power and utility *without* being reused as formal certification evidence. Separate design and screen splits are an auditable implementation choice, not a minimal validity requirement: they may be merged or cross-fitted as long as the complete family and order are frozen before D_{cert} is inspected.

$$v_\theta^\lambda \in [-\lambda, 1]. \tag{C.3}$$

Once model answers and uncertainty scores are cached, all routing actions for a candidate θ_k are deterministic. A direct implementation evaluates K policies on m certification units in $O(Km)$ time, after candidate construction and screening. The exact binomial p-value can be computed using a standard binomial CDF in log-stable numerical libraries. Because fixed-sequence testing stops at the first failure, the number of p-values actually needed can be substantially smaller than K, although utilities for all certified policies are still evaluated to choose Eq. (11).

For a quantile grid with K_P Primary thresholds and K_G Guardian thresholds, the unpruned family has at most $K_P K_G$ candidates. Design-stage action-vector de-duplication can reduce this number without affecting validity, because it occurs before certification and simply defines the finite family that will later be tested.

D Relationship to the Previous Stability-Correction Route

An earlier route to non-monotonic risk control would optimize empirical utility under a tightened risk constraint of the form

$$\widehat{R}_n(\theta) \le \alpha - b_n, \tag{D.1}$$

where b_n must upper-bound an optimizer-stability error. Such an approach can be elegant when a positive and deployable stability margin is available, and it is connected to the general non-monotonic CRC framework (Angelopoulos, 2026). In the present two-threshold routing problem, however, flat objective regions, near-ties, offsetting positive and negative risk changes, and action aliases make the required stability constants difficult to certify robustly from finite data.

FS-Cert-U removes b_n from the deployed algorithm. Its guarantee is instead built from two verifiable objects: a finite family frozen before certification and valid candidate-wise p-values computed on independent certification data. The proof obligations are therefore different and should not be mixed: stability-based guarantees certify a data-dependent optimizer through its sensitivity, whereas FS-Cert-U certifies the policies themselves before allowing utility optimization.

E Additional Experimental Details and Results

E.1 Benchmarks, models, and implementation

Dataset construction and correctness. CommonsenseQA uses a fixed subset of 10,000 questions. TriviaQA starts from 8,000 sampled occurrences and retains one occurrence per question group, yielding 6,416 distinct questions. ScienceQA uses the 2,017 image-containing questions in its test set. Multiple-choice correctness is agreement with the reference option. TriviaQA correctness is exact match after lowercasing, replacing underscores with spaces, removing ASCII punctuation and English articles, and normalizing whitespace; all reference aliases are allowed. Empty or truncated generations remain in the evaluation. MM-Vet v2 excludes one question with an empty reference answer from the 517-question cache, leaving 516 questions. Its scoring protocol is specified in Appendix E.7.

Table 2: Benchmark and partition sizes. K is the number of distinct routing candidates with the default five threshold levels.

Dataset	Total	Design	Screen	Cert.	Test	K
CommonsenseQA	10,000	1,000	2,000	3,500	3,500	21
TriviaQA	6,416	641	1,283	2,246	2,246	21
ScienceQA	2,017	201	404	706	706	7–17
MM-Vet v2	516	51	103	181	181	21

Model pairs. Table 3 gives the directed pairs. All model pairs below use full model names; Qwen3.5, Qwen3.8, Gemma 4, Ministral 3, and Nemotron 3 identify the model families. Nemotron 3-30B-A3B is the Nano Omni reasoning model. All ten models are used on the three main datasets. The five prespecified representative pairs used in main-text pair-level figures are Gemma 4-12B → Gemma 4-31B, Ministral 3-8B → Ministral 3-14B, Qwen3.5-2B → Qwen3.5-27B, Qwen3.5-9B → Nemotron 3-30B-A3B, and Qwen3.5-9B → Qwen3.5-27B; they span within-family and cross-family routing, capacity gaps, and Primary strengths. The Qwen3.5-2B → Qwen3.5-9B pair is additionally used for the detailed allocation analysis because cached routing records are available. MM-Vet uses nine models; its two pairs involving Nemotron are replaced by Ministral 3-8B → Gemma 4-31B and Qwen3.5-9B → Gemma 4-31B.

Table 3: Model pairs evaluated on each benchmark. QA denotes CommonsenseQA and TriviaQA.

Primary	Guardian	QA / ScienceQA	MM-Vet v2
Qwen3.5-2B	Qwen3.5-9B	✓	✓
Qwen3.5-2B	Qwen3.5-27B	✓	✓
Qwen3.5-9B	Qwen3.5-27B	✓	✓
Gemma 4-E4B	Gemma 4-12B	✓	✓
Gemma 4-E4B	Gemma 4-31B	✓	✓
Gemma 4-12B	Gemma 4-31B	✓	✓
Ministral 3-8B	Ministral 3-14B	✓	✓
Qwen3.5-2B	Gemma 4-12B	✓	✓
Ministral 3-8B	Nemotron 3-30B-A3B	✓	–
Qwen3.5-9B	Nemotron 3-30B-A3B	✓	–
Qwen3.5-9B	Qwen3.8-27B	✓	✓
Ministral 3-8B	Gemma 4-31B	–	✓
Qwen3.5-9B	Gemma 4-31B	–	✓

Generation and uncertainty. CommonsenseQA uses native chat templates with a fixed one-shot example, chooses the highest-scoring answer option, and uses entropy over option probabilities as uncertainty. TriviaQA uses a fixed one-shot prompt with greedy generation and a 36-token limit; uncertainty is the mean entropy of the full-vocabulary token distributions along the generated answer. ScienceQA uses native image input without answer hints, chooses the option with the largest next-token logit, and computes uncertainty from the empirical frequencies of 20 option samples at temperature 1. MM-Vet uses native multi-image input, greedy generation with a 32-token limit, and mean token entropy. All routing methods operate on the same cached predictions and uncertainty values.

Candidate construction and ordering. For L threshold levels, each model contributes the design-score quantiles at fractions $j/(L-1)$, $j = 1, \ldots, L-2$, together with $-\infty$ and $+\infty$. Quantiles use the higher observed order statistic. Duplicate threshold values are removed. The Cartesian product is then reduced by retaining only one representative with $\tau_P = +\infty$, because the Guardian threshold is irrelevant for that policy on

every input. Thus the maximum family size is $L(L-1)+1$: 7, 21, and 73 for $L = 3, 5, 9$. Tied uncertainty scores can further reduce K, as on ScienceQA. The screen order sorts by increasing exact risk p-value, decreasing empirical utility, and candidate index. Each tested target α has its own screen order, frozen before certification. The utility maximizer breaks ties by candidate index.

Compared implementations. The main results use the positive-coverage risk certificate in Eq. (8). The optional coverage-floor extension is evaluated separately in Appendix E.5. Bonferroni evaluates all K candidates at level δ/K; Holm uses its own p-value ordering and step-down cutoffs; FS-Cert-U stops at the first failure at level δ. These procedures select by certification utility. Naive uses the empirical constraint $\widehat{\mathbb{E}}[Z] \le \alpha\widehat{\mathbb{E}}[S]$ and selects the feasible candidate with highest empirical utility. Primary-FS and Guardian-FS use the distinct marginal thresholds from the routing grid, their own screen order, and the same certification data. Their utility selection reduces to maximizing coverage because their Guardian-call costs are constant. The recorded routing fallback calls the Guardian and then abstains, so its utility is $-\lambda$; Primary-FS's all-abstain outcome costs zero. The same accounting applies to every split in the reported results.

Matched splits. Question groups are randomly permuted using seeds 929000–929499 for the three main datasets and 961000–961499 for MM-Vet. Partition boundaries are $\lfloor 0.1N \rfloor$, $\lfloor 0.3N \rfloor$, and $\lfloor 0.65N \rfloor$ in the ordered groups. Table 2 reports the resulting sample sizes. The same split is reused across methods and targets within each model-pair experiment. The sensitivity study on the three main datasets uses the first 50 of these seeds for every configuration, including the default configuration.

E.2 Metric definitions and aggregation

For model pair j and split b, let n_{jb} be the test size, M_{jb} the number answered, E_{jb} the number of accepted errors, and C_{jb} the number of Guardian calls. We compute

$$\widehat{q}_{jb} = \frac{E_{jb}}{M_{jb}} \quad (M_{jb} > 0), \qquad \widehat{s}_{jb} = \frac{M_{jb}}{n_{jb}}, \qquad \widehat{U}_{jb} = \frac{M_{jb} - \lambda C_{jb}}{n_{jb}}. \tag{E.1}$$

Accepted correct count is $M_{jb} - E_{jb}$. Primary and Guardian allocations divide their respective accepted counts by n_{jb}; they sum to coverage. Guardian calls include both accepted Guardian answers and subsequent abstentions.

Risk curves average $\widehat{q}_{jb}$ over the splits with $M_{jb} > 0$ for each model pair, with bands showing their sample standard deviation. The SCER entries in Tables 4 and 11 instead first pool counts within a pair,

$$\widetilde{q}_j = \frac{\sum_b E_{jb}}{\sum_b M_{jb}}, \tag{E.2}$$

then average $\widetilde{q}_j$ over pairs with a nonzero denominator. Coverage, utility, calls, and accepted correct counts include all splits. In dataset-level utility plots, each point first averages the model pairs within each split, and the displayed mean and standard deviation summarize those 500 averages.

The test exceedance rate is

$$\frac{\sum_{j,b} \mathbf{1}\{M_{jb} > 0,\ E_{jb}/M_{jb} > \alpha\}}{\sum_{j,b} \mathbf{1}\{M_{jb} > 0\}}. \tag{E.3}$$

It summarizes finite-test-sample behavior; the event differs from certifying a population-unsafe policy in Theorem 1. The 500 repetitions resplit a fixed benchmark, and standard deviations describe that split variability. The intervals in Figure 24 are the 2.5th–97.5th percentiles of paired split differences.

For the recovery analysis, the denominator is the number of test questions for which at least one model has a correct cached answer. Dividing accepted correct count by this common denominator measures the fraction of available correct answers recovered by routing. It differs from single-model power, whose denominator includes only correct answers from that model. Each ratio is computed within a split before averaging.

Expanded primary-operating-point comparison. For completeness, Table 4 reports the expanded version of the primary-operating-point comparison, including the separately certified Primary-FS and Guardian-FS references, nonempty test-output rates, and accepted-correct counts omitted from the compact main-text table.

Table 4: Expanded comparison at $\alpha = 0.10$, $\lambda = 0.25$. Equal-weight means over all 11 pairs per dataset and 500 matched splits. SCER pools counts within each pair; nonempty refers to test answers. Correct is the number of accepted correct test answers.

Method	SCER (%)	Cov. (%)	Guardian calls (%)	Utility	Nonempty (%)	Correct
CommonsenseQA						
Primary-FS	6.87	28.53	0.00	0.285	73.00	946.9
Guardian-FS	6.29	56.57	100.00	0.316	100.00	1850.2
Naive	8.66	64.84	62.14	0.493	100.00	2073.6
Bonferroni	6.73	59.32	72.89	0.411	100.00	1933.3
Holm	6.77	59.51	72.55	0.414	100.00	1938.6
FS-Cert-U	7.41	61.41	69.11	0.441	100.00	1987.1
TriviaQA						
Primary-FS	6.40	14.81	0.00	0.148	59.69	311.9
Guardian-FS	6.31	24.57	100.00	-0.004	88.82	517.8
Naive	8.32	39.09	85.82	0.176	99.95	803.3
Bonferroni	6.07	23.24	89.66	0.008	81.36	491.2
Holm	6.09	23.31	89.61	0.009	81.36	492.5
FS-Cert-U	6.86	29.40	86.88	0.077	93.00	615.0
ScienceQA						
Primary-FS	6.66	32.83	0.00	0.328	54.80	222.9
Guardian-FS	6.75	65.14	100.00	0.401	73.36	434.1
Naive	9.03	84.95	53.30	0.716	94.82	546.6
Bonferroni	6.08	59.27	68.86	0.421	69.35	395.1
Holm	6.14	59.39	68.63	0.422	69.35	395.5
FS-Cert-U	6.80	69.75	67.11	0.530	81.58	461.6

E.3 Risk, utility, and allocation across configurations

Figures 11–13 include all 33 main model pairs. Figure 14 complements these mean-risk curves with the test exceedance rate and zero-output frequency. Figure 24 shows the paired utility difference at $\alpha = 0.10$ for every pair. Figures 15 and 16 compare routing with separately certified models in terms of allocation and recovery of correct answers.

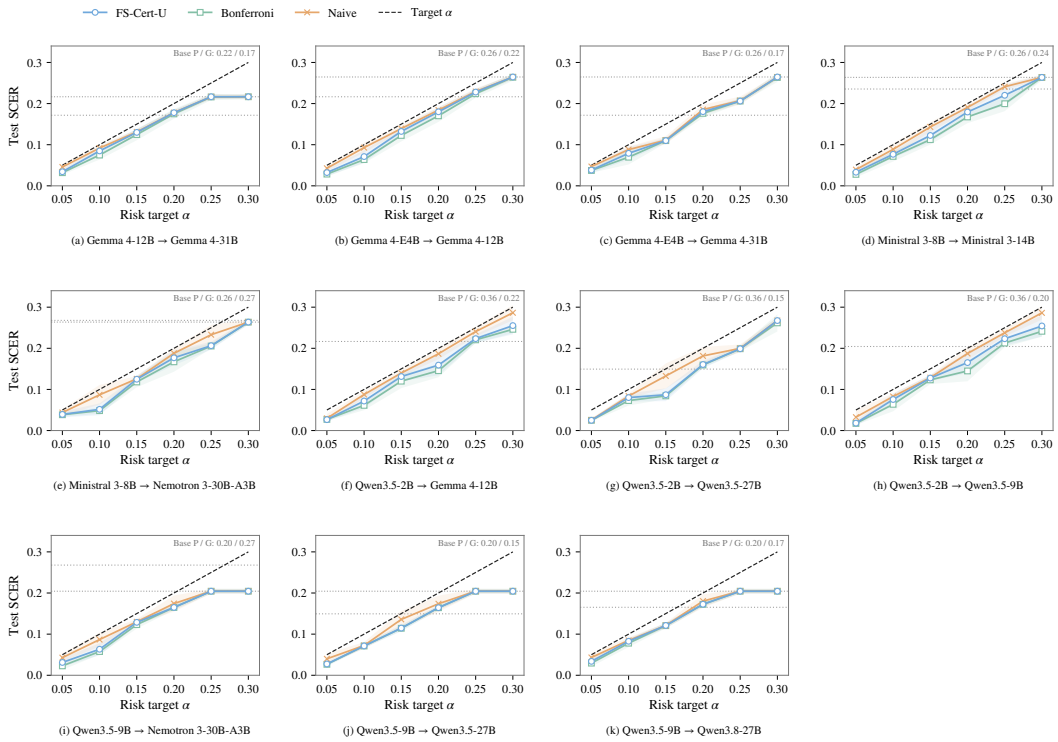


Figure 11: CommonsenseQA: test SCER for all 11 model pairs (mean ± standard deviation over nonempty splits). Gray references give the unconditional models' error rates; dashed diagonals indicate α.

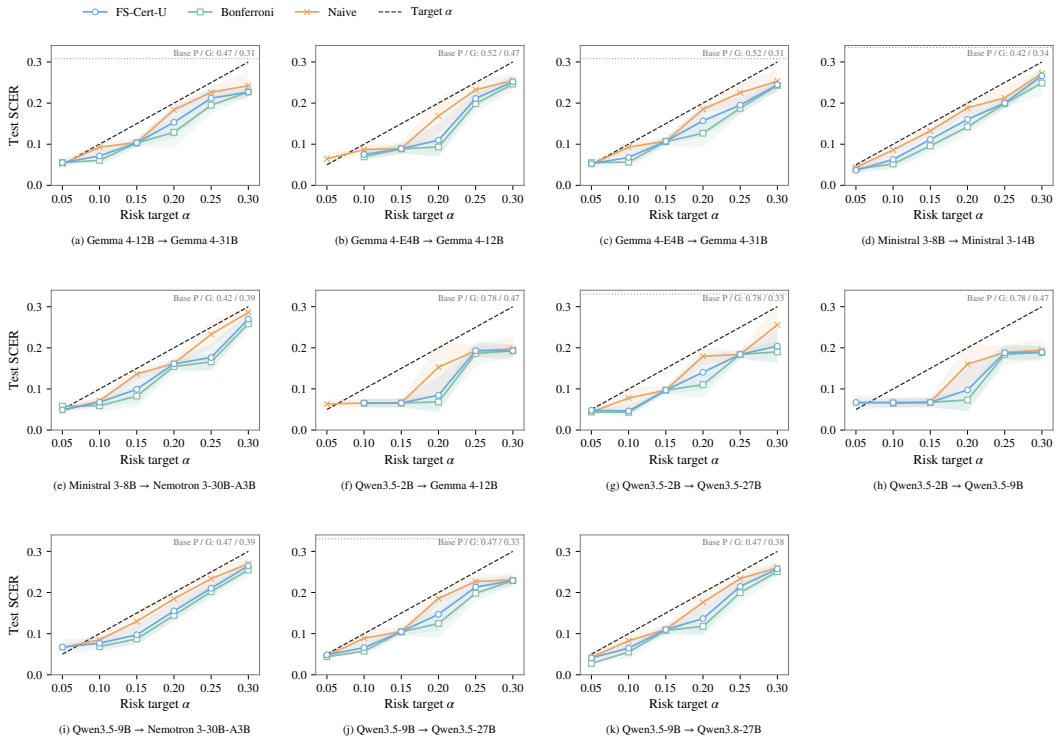


Figure 12: TriviaQA: test SCER for all 11 model pairs. The aggregation and line conventions follow Figure 11.

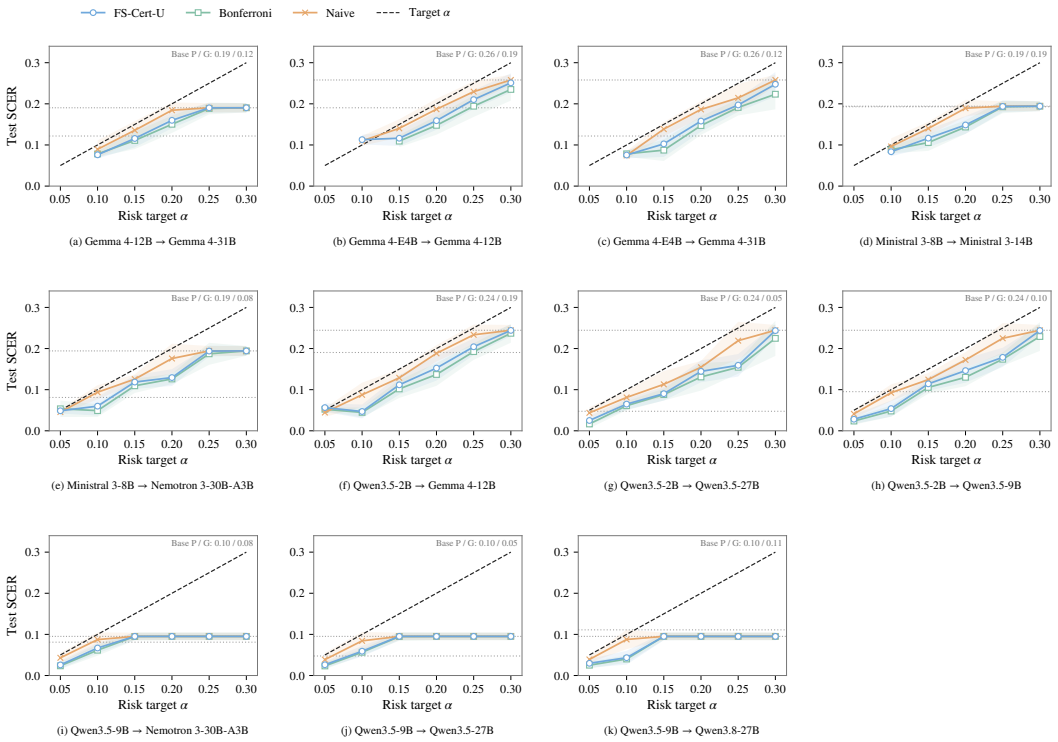


Figure 13: ScienceQA: test SCER for all 11 model pairs. Targets with no answered test examples have no risk point; zero-output frequencies are reported in Figure 14.

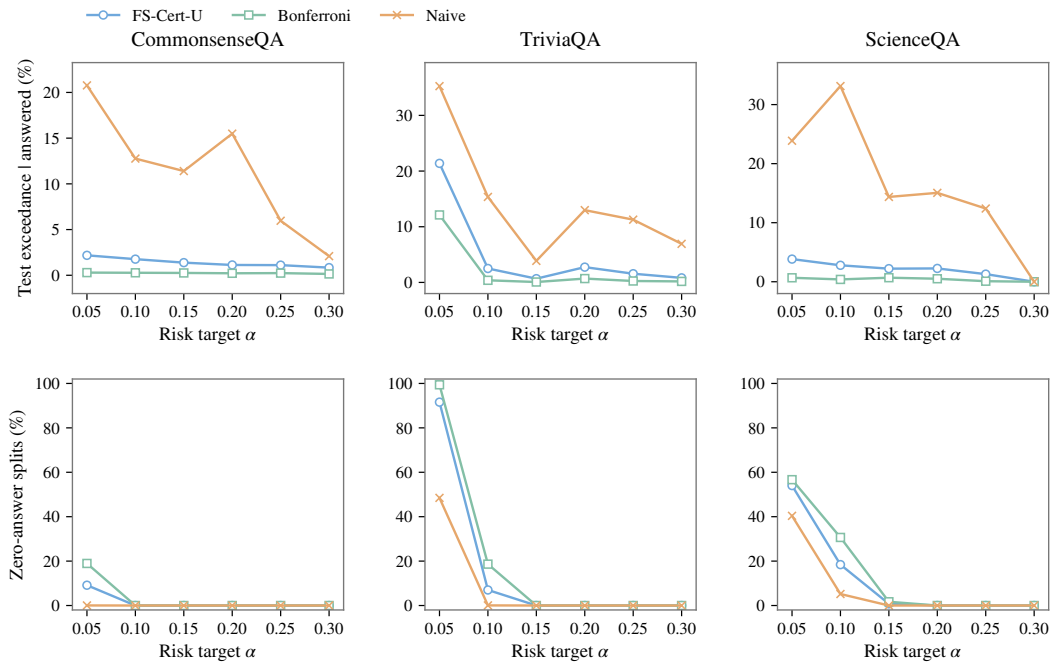


Figure 14: Test exceedance among nonempty splits (top) and zero-output frequency among all splits (bottom), pooled across the 11 model pairs per dataset.

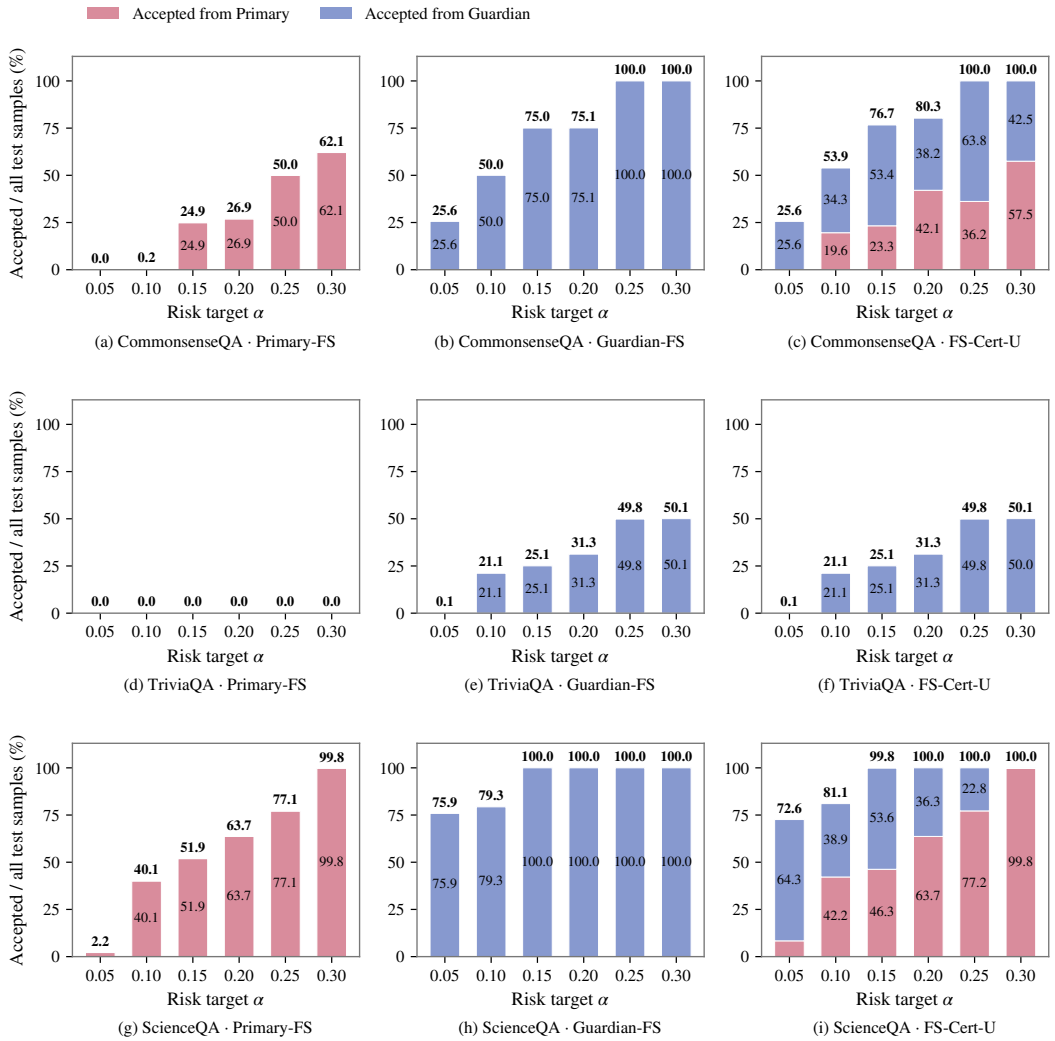


Figure 15: Separately certified models and routing for Qwen3.5-2B → Qwen3.5-9B. Segments show the fraction of test queries answered by each model, and their sum is coverage (means over 500 matched splits).

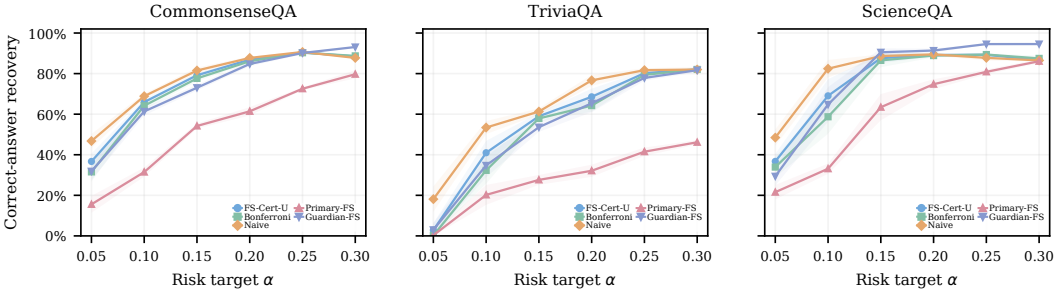


Figure 16: Correct-answer recovery using the common denominator of test questions answered correctly by at least one model. Means and standard deviations summarize 500 split-wise averages over model pairs.

E.4 SENSITIVITY AND MECHANISM STUDIES

Calibration data and grid resolution. Figure 17 varies the certification fraction $f \in \{0.25, 0.5, 1\}$ while retaining the same design, screen, and test partitions. Unused certification examples are not transferred to another split. It also varies threshold levels $L \in \{3, 5, 9\}$ and cost $\lambda \in \{0, 0.1, 0.25, 0.5\}$, one factor at a time. All panels use $\alpha = 0.10$. Increasing the certification fraction improves FS-Cert-U utility from 0.2241 to 0.2995 and 0.3459. A finer threshold grid improves utility from 0.1837 to 0.3459 and 0.3951. The mean utility advantage over Bonferroni is 0.0597, 0.0627, 0.0674, and 0.0750 for the four cost weights, respectively.

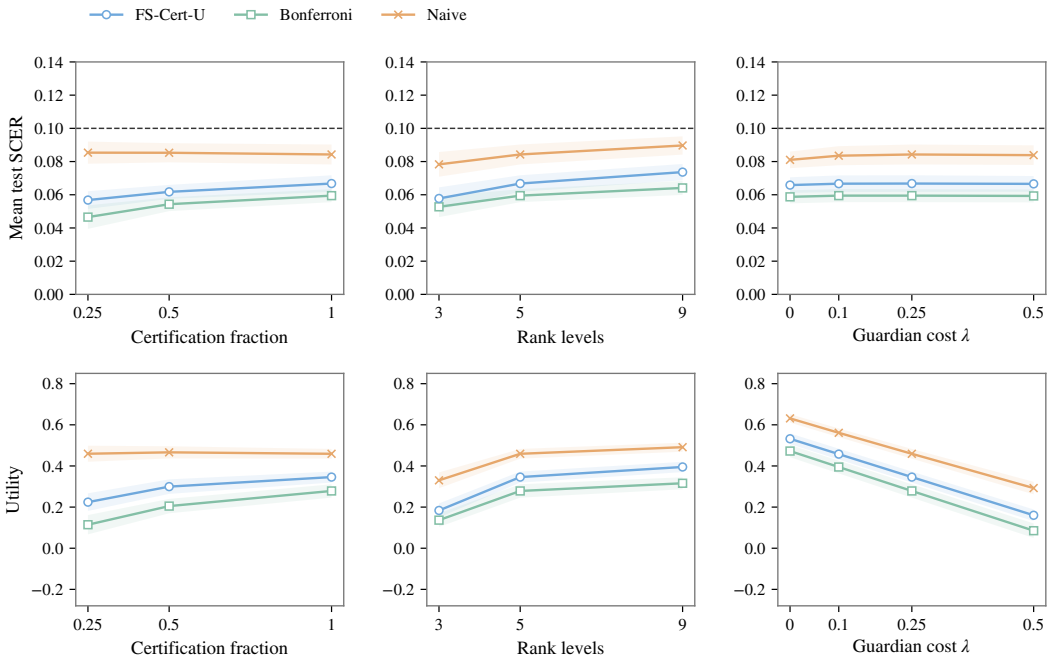


Figure 17: Sensitivity to certification fraction, threshold levels, and Guardian cost at $\alpha = 0.10$. Points summarize 50 matched splits over all 33 main configurations; bands indicate one standard deviation. The default point also uses these 50 splits.

Certification versus final selection. An additional six-configuration study compares selection by screen utility with selection by certification utility inside the same fixed-sequence set. This study uses the earlier screen-based candidate-construction protocol, so it isolates the certification and selection mechanisms separately from the main design-quantile experiment. Table 5 shows mean gains of 0.0269–0.0848 for fixed-sequence versus Bonferroni certification. Changing the within-set utility estimator has a much smaller effect, with absolute mean differences below 1.5×10^{-4}. In these configurations, the main empirical gain comes from the certification procedure; the utility theorem establishes that subsequent optimization can reuse certification data safely.

The study assigns 50% of questions to testing and divides the remaining half 30%/70% between screening and certification. It constructs up to 64 candidates from the screen-set utility frontier across risk budgets between zero and α, then freezes that family and its order before certification. Within each split, the two selectors share exactly the same accepted set; only the utility estimates used for final selection differ.

Table 5: Certification and selector ablations at $\alpha = 0.10$ (500 paired splits). The second comparison changes only the utility estimator used to select within the same certified set.

Configuration	$\Delta U_{\mathrm{FS-Bonf}}$	$\Delta U_{\mathrm{cert-screen}}$
CommonsenseQA / Gemma 4-E4B → Gemma 4-12B	0.033359	-0.000058
CommonsenseQA / Ministral 3-8B → Ministral 3-14B	0.035880	0.000041
CommonsenseQA / Qwen3.5-2B → Qwen3.5-9B	0.026930	-0.000043
ScienceQA / Qwen3.5-2B → Qwen3.5-9B	0.043280	0.000148
TriviaQA / Ministral 3-8B → Ministral 3-14B	0.084777	-0.000012
TriviaQA / Qwen3.5-2B → Qwen3.5-9B	0.063346	0.000011

E.5 COMPACT SUPPLEMENTARY MECHANISM RESULTS

The following replay uses the same frozen predictions, candidate IDs, split hashes, utility definition, and tie-breaking rules as the main experiment. It covers all 44 configurations (33 main QA configurations and 11 MM-Vet v2 proxy configurations), with 500 matched splits per configuration and six risk targets unless otherwise noted. The compact figure reports the three main datasets at $\alpha = 0.10$; the accompanying tables average equally over their 33 model-pair configurations. All complete model-pair, target, candidate-level, and matched-difference records are retained in the supplementary artifact directories.

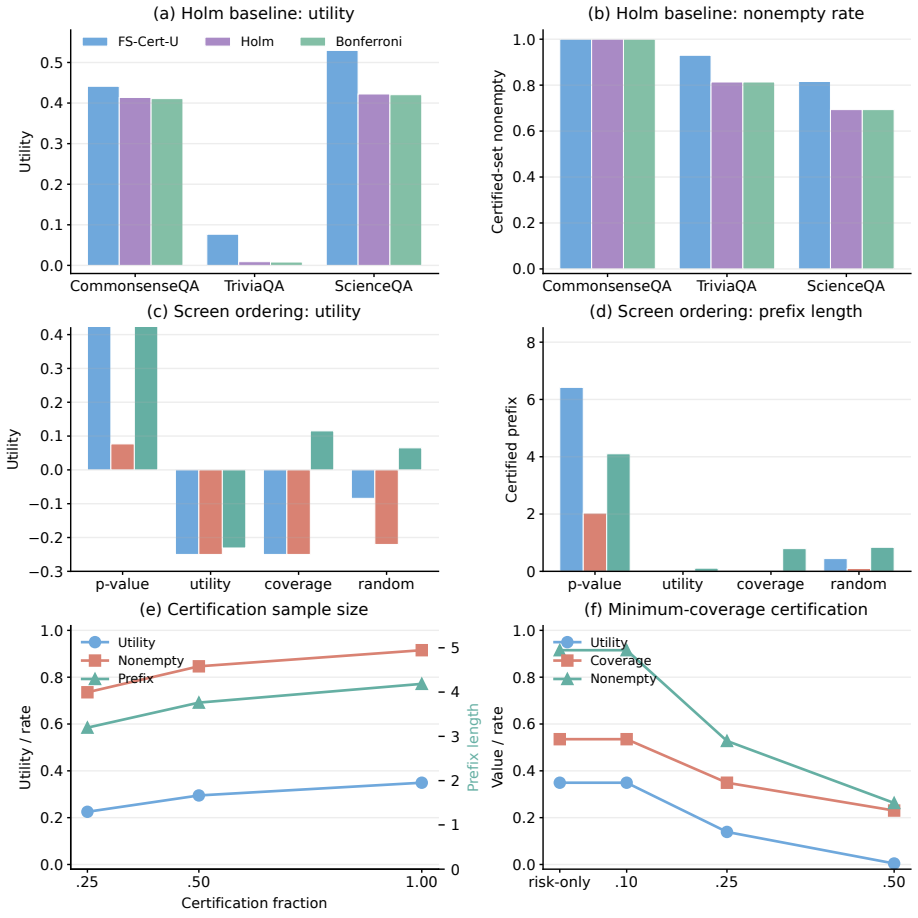


Figure 18: **Compact mechanism summary on the 33 main configurations.** Panels (a–b) compare FS-Cert-U, Holm, and Bonferroni; (c–d) show how the frozen screen order changes utility and certified-prefix length; (e) varies certification sample size; and (f) adds a joint minimum-coverage requirement. Bars and points are equal-weight means over model pairs at $\alpha = 0.10$; the underlying split-level distributions and all 44 configurations are archived with the paper artifacts.

Table 6: Compact supplementary comparison at $\alpha = 0.10$. Utility and certification nonempty rate are equal-weight means over the 33 main model-pair configurations. The final block reports matched FS–single-model utility differences and win/tie/loss counts.

Dataset	FS U	Holm U	Bonf. U	FS nonempty	Holm nonempty	Bonf. nonempty
CommonsenseQA	0.441	0.414	0.411	1.000	1.000	1.000
TriviaQA	0.077	0.009	0.008	0.930	0.814	0.814
ScienceQA	0.530	0.422	0.421	0.816	0.693	0.693

Baseline	ΔU (FS − baseline)	Wins	Ties	Losses
Primary-FS	+0.095	19	0	14
Guardian-FS	+0.112	25	4	4

Table 7: Mechanism ablations, averaged over the 33 main configurations at $\alpha = 0.10$. For ordering and sample-size rows, the fourth column is certification nonempty rate; for minimum-coverage rows it reports test coverage / certification nonempty rate.

Experiment	Setting	Utility	Coverage / nonempty	Prefix
Ordering	p-value first	0.349	0.915	4.19
Ordering	utility first	-0.244	0.005	0.04
Ordering	coverage first	-0.128	0.107	0.26
Ordering	random	-0.080	0.232	0.46
Cert. fraction	0.25	0.225	0.736	3.20
Cert. fraction	0.50	0.295	0.846	3.76
Cert. fraction	1.00	0.349	0.915	4.19
Min. coverage	risk-only	0.349	0.535 / 0.915	4.19
Min. coverage	0.10	0.349	0.535 / 0.915	4.19
Min. coverage	0.25	0.139	0.349 / 0.528	2.05
Min. coverage	0.50	0.004	0.230 / 0.263	1.23

Scenario	Method	FWER ($\delta = .05$)	Pop. regret	Test regret
all unsafe boundary	Bonferroni	0.012	0.0000	0.0000
all unsafe boundary	FS-Cert-U	0.031	0.0000	0.0000
all unsafe boundary	Holm	0.012	0.0000	0.0000
six type original	Bonferroni	0.005	0.0000	0.0000
six type original	FS-Cert-U	0.022	0.0000	0.0000
six type original	Holm	0.008	0.0000	0.0000
strong conflict	Bonferroni	0.000	0.0000	0.0000
strong conflict	FS-Cert-U	0.001	0.0000	0.0000
strong conflict	Holm	0.000	0.0000	0.0000

Table 8: Compact summary of direct FWER and utility-regret validation at $m = 4480$. Regret columns average over the displayed δ values.

E.6 Direct validation of FWER and utility regret

We run a controlled-population replay with a large oracle pool to measure the two theorem events directly. The frozen family contains 27 routing candidates and each scenario is a six-type mixture. For each of the three error scenarios (the original mixture, a stronger Primary–Guardian conflict mixture, and an all-unsafe boundary mixture), we draw an independent oracle pool of two million examples and define a candidate as oracle-unsafe when its large-pool SCER estimate q_k^{oracle} is at least $\alpha = 0.10$. We then draw 5,000 independent screen, certification, and test splits for each $\delta \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$ and certification size $m \in \{1120, 4480\}$. The oracle pool is an approximation to population truth; its maximum SCER error relative to exact type-level enumeration is 7.2×10^{-4}.

For each split, the FWER indicator is one when any unsafe candidate is certified. Figure 26 in the main text plots its empirical mean with the diagonal $y = \delta$ reference. The rates remain below the target across the three mixtures and all methods, with the expected increase as δ grows. For utility, we compute both the population regret and the independent-test regret relative to the best accepted candidate. The dashed line uses $\eta = 0.05$ and the displayed $m = 4480$ in $2\varepsilon_U$. Regret is exactly zero for every nonempty certified set in this replay; although the bound holds, these populations do not test its sharpness. Empty certified sets are retained as missing regret values rather than treated as zero. All split hashes, candidate-level decisions, and verification checks are included in the supplementary experiment directory.

Interpretation. At $\alpha = 0.10$, FS-Cert-U has mean utility 0.349 across the main configurations, compared with 0.282 for Holm and 0.280 for Bonferroni. The p-value-first order yields a mean certified prefix of 4.19 candidates; utility-first, coverage-first, and random orders yield 0.04, 0.26, and 0.46, respectively. Increasing the certification fraction from 0.25 to 1 increases utility from 0.225 to 0.349. The joint coverage floor exhibits the expected trade-off: $s_{\min} = 0.50$ sharply lowers both the certified-set nonempty rate and utility, while excluding low-coverage certified policies.

Known-population certification. The earlier synthetic certification study is separate from the replay above. It uses the same three discrete populations but only 300 fresh i.i.d. trials per setting, total calibration size $n_{\mathrm{cal}} \in \{400, 1600\}$, and target $\alpha \in \{0.05, 0.10, 0.20\}$. Calibration data are split 30%/70% between screening and certification, so its certification size is $m \in \{280, 1120\}$. Because the population risks are known, the

measured failure event is that at least one candidate with $q_k \geq \alpha$ enters the certified set. Table 9 reports the counts. Fixed-sequence failure frequencies range from 0 to 3.33% across these settings at $\delta = 0.05$; Bonferroni frequencies range from 0 to 2.00%.

The six input types are equally likely, with uncertainty scores $u_P = (0.05, 0.15, 0.30, 0.50, 0.70, 0.90)$ and $u_G = (0.80, 0.20, 0.60, 0.10, 0.90, 0.30)$. Conditional on input type, Primary and Guardian errors are independent Bernoulli draws with probabilities

$$\begin{aligned} e_P^{\text{mixture}} &= (0.02, 0.04, 0.25, 0.12, 0.45, 0.65), \\ e_G^{\text{mixture}} &= (0.22, 0.02, 0.03, 0.18, 0.08, 0.02), \\ e_P^{\text{conflict}} &= (0.015, 0.16, 0.035, 0.32, 0.075, 0.48), \\ e_G^{\text{conflict}} &= (0.30, 0.025, 0.26, 0.035, 0.19, 0.055). \end{aligned} \tag{E.4}$$

In the boundary population, both models have error probability $\alpha + 0.001$ for every type. Enumerating thresholds at the six observed score values and $-\infty$, removing duplicate action vectors, and excluding all-abstain gives a fixed family of 27 candidates. Population coverage and risk are computed exactly by weighting the six types; calibration samples are drawn independently in each trial.

Table 9: Certification failures on known synthetic populations: a failure means that at least one population-unsafe candidate is certified. Each cell uses 300 fresh i.i.d. trials and $\delta = 0.05$. Calibration size n_{cal} is split 30%/70% between screening and certification.

Population	n_{cal}	α	FS failures	Bonf. failures
Unsafe boundary	400	0.05	3/300	2/300
Unsafe boundary	400	0.10	8/300	5/300
Unsafe boundary	400	0.20	6/300	1/300
Unsafe boundary	1600	0.05	8/300	3/300
Unsafe boundary	1600	0.10	7/300	6/300
Unsafe boundary	1600	0.20	10/300	5/300
Six-type mixture	400	0.05	1/300	0/300
Six-type mixture	400	0.10	2/300	2/300
Six-type mixture	400	0.20	0/300	0/300
Six-type mixture	1600	0.05	0/300	0/300
Six-type mixture	1600	0.10	1/300	0/300
Six-type mixture	1600	0.20	0/300	0/300
Strong conflict	400	0.05	0/300	0/300
Strong conflict	400	0.10	1/300	0/300
Strong conflict	400	0.20	0/300	0/300
Strong conflict	1600	0.05	0/300	0/300
Strong conflict	1600	0.10	3/300	0/300
Strong conflict	1600	0.20	0/300	0/300

E.7 MM-VET V2: SIMILARITY-BASED CORRECTNESS

MM-Vet v2 is a fourth-domain stress test rather than a benchmark directly comparable to the three primary QA datasets: correctness uses proxy sentence similarity and the certification set is materially smaller. Table 10 reports selected risk targets across all 11 pairs; the complete target sweep and pair-level diagnostics follow.

Scoring. For each generated answer, we evaluate the sentence-pair model `cross-encoder/stsb-roberta-large`[1] on the pair (question + reference, question + prediction). An answer is labeled correct when the sigmoid of the model's logit is at least 0.6. This is the similarity-based evaluation criterion; it is distinct from the official MM-Vet scoring protocol. All methods use the same labels, cached generations, and 500 question-level splits.

Table 11 reports all six risk targets. At $\alpha = 0.10$, FS-Cert-U answers at least one test question in 107 of 5,500 pair–split evaluations. Its pair-pooled macro SCER is 12.69%, and 48 of those 107 nonempty splits exceed the target. The very low coverage therefore accompanies considerable test-risk variability. At $\alpha = 0.25$, coverage increases to 46.89%, with SCER of 18.36% and utility of 0.2719. Figure 19 shows the individual model pairs.

Candidate-density study. A separate matched replay varies the grid from 5 to 33 threshold levels, increasing median K from 21 to 1,057. Table 12 shows that denser grids alone do not resolve the low-output regime.

[1] `https://huggingface.co/cross-encoder/stsb-roberta-large`

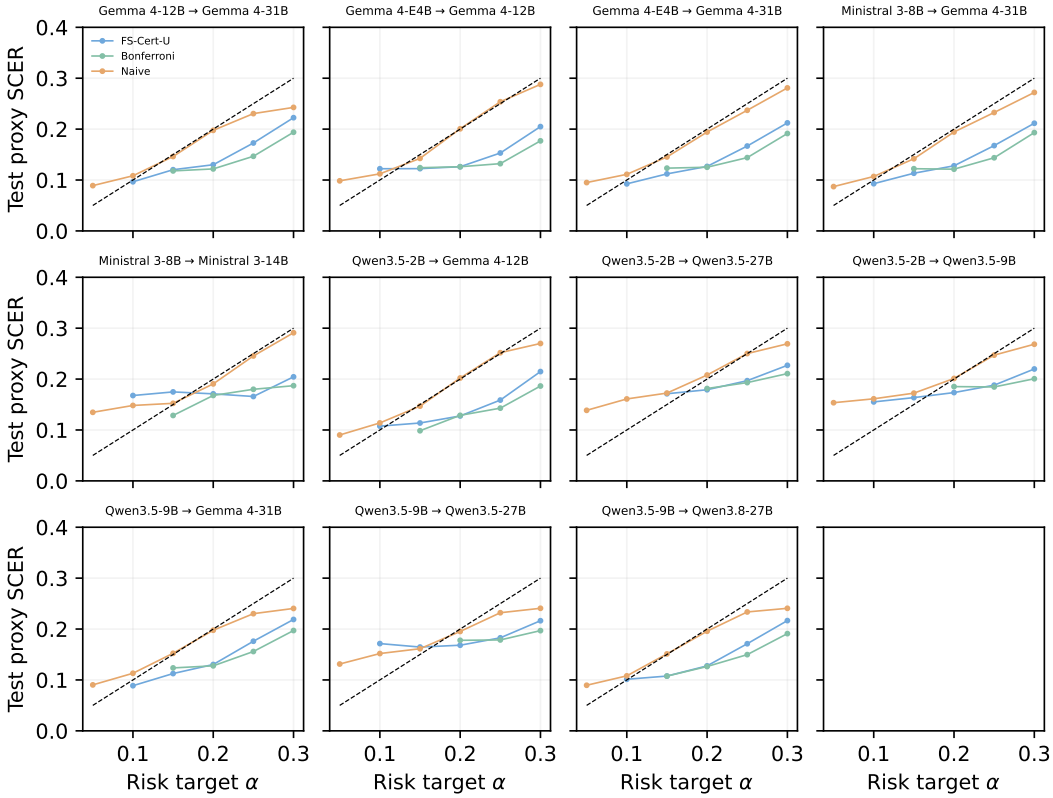


Figure 19: MM-Vet v2: similarity-label SCER for all 11 pairs (mean ± standard deviation over nonempty splits). Panels report the number of nonempty FS-Cert-U splits at $\alpha = 0.10$.

Table 10: MM-Vet v2 stress test at selected risk targets. Correctness uses proxy sentence similarity and the certification set is small; results are contextual rather than directly comparable to the three QA benchmarks.

α	Method	SCER (%)	Coverage (%)	Utility	Nonempty (%)
0.10	Bonferroni	–	0.00	-0.2500	0.00
	Naive	13.53	23.22	0.0018	63.71
	FS-Cert-U	12.69	0.44	-0.2455	1.95
0.25	Bonferroni	16.90	32.26	0.1036	57.33
	Naive	24.31	93.69	0.8532	100.00
	FS-Cert-U	18.36	46.89	0.2719	71.80
0.30	Bonferroni	20.23	68.23	0.5137	88.33
	Naive	26.44	99.44	0.9713	100.00
	FS-Cert-U	22.35	76.39	0.6333	90.73

Table 11: MM-Vet v2 with sentence-similarity correctness labels. SCER uses the same per-pair pooling as Table 4; nonempty reports the percentage of test splits with at least one answer. The reported error rates concern the similarity criterion.

α	Method	SCER (%)	Coverage (%)	Utility	Nonempty (%)
0.05	Guardian-FS	–	0.00	-0.2500	0.00
	Bonferroni	–	0.00	-0.2500	0.00
	Naive	11.60	5.42	-0.1928	21.33
	FS-Cert-U	–	0.00	-0.2500	0.00
0.10	Guardian-FS	11.99	0.64	-0.2436	2.80
	Bonferroni	–	0.00	-0.2500	0.00
	Naive	13.53	23.22	0.0018	63.71
	FS-Cert-U	12.69	0.44	-0.2455	1.95
0.15	Guardian-FS	13.43	6.37	-0.1863	17.56
	Bonferroni	12.39	0.79	-0.2418	2.22
	Naive	16.23	49.04	0.2985	92.51
	FS-Cert-U	14.39	4.85	-0.1989	13.73
0.20	Guardian-FS	15.04	23.47	-0.0153	47.98
	Bonferroni	15.17	8.31	-0.1614	19.71
	Naive	20.44	77.32	0.6253	99.69
	FS-Cert-U	15.34	18.77	-0.0471	40.71
0.25	Guardian-FS	18.38	54.90	0.2990	76.47
	Bonferroni	16.90	32.26	0.1036	57.33
	Naive	24.31	93.69	0.8532	100.00
	FS-Cert-U	18.36	46.89	0.2719	71.80
0.30	Guardian-FS	21.11	82.84	0.5784	92.80
	Bonferroni	20.23	68.23	0.5137	88.33
	Naive	26.44	99.44	0.9713	100.00
	FS-Cert-U	22.35	76.39	0.6333	90.73

At $\alpha = 0.10$, coverage peaks at 0.686% with nine levels and returns to 0.488% with 33 levels. Although the fraction of splits with some candidate passing increases from 4.58% to 20.93%, many such candidates occur after the first failed test and are excluded by fixed-sequence stopping. At $\alpha = 0.25$, the output rate remains near 72%, while utility decreases from 0.2720 to 0.2263. Candidate resolution, ordering, and certification sample size therefore play distinct roles in determining the useful certified prefix.

Table 12: MM-Vet v2 candidate-density study (500 splits, 11 pairs). Output, coverage and utility use the fixed-sequence prefix. Any pass counts splits with at least one candidate p-value $\leq \delta$, including candidates beyond the first failure.

Levels	Median K	Output (%)	Coverage (%)	Utility	Any pass (%)
$\alpha = 0.10$					
5	21	1.95	0.435	-0.2455	4.58
9	73	3.33	0.686	-0.2428	11.25
17	273	2.35	0.498	-0.2447	17.36
33	1057	2.24	0.488	-0.2447	20.93
$\alpha = 0.25$					
5	21	71.80	46.885	0.2720	93.36
9	73	71.98	45.549	0.2607	97.42
17	273	72.07	43.545	0.2402	98.67
33	1057	72.07	42.287	0.2263	99.13

E.8 External Comparison with LEC-Routing

This comparison contextualizes FS-Cert-U against a prior risk-calibrated two-model router, rather than isolating the effect of the certification rule. **LEC obtains higher average coverage and utility than FS-Cert-U on all three representative-pair dataset averages.** The methods optimize over different policy spaces and target different risk statements, so these results are not a controlled fixed-sequence-versus-LEC ablation.

LEC-Routing maximizes calibration retention under the published constraint $\sum_i (Z_i - \alpha S_i) \leq -1$, following Eqs. (9)–(11) of Wang et al. (2026a). It uses the full effective calibration-score grid, whereas FS-Cert-U uses a finite family frozen before certification and provides family-wise protection for the certified policy set exposed to downstream utility optimization, together with a separate utility-regret guarantee. LEC's marginal-risk statement differs from this inferential target; we do not rank the guarantees without qualification. We use neither an extra risk margin nor the archived accepted-correct objective.

Table 13 uses only the five displayed pairs for every method, unlike the all-11-pair summary in Table 4. LEC obtains higher coverage and utility than FS-Cert-U on all three dataset averages: at $\alpha = 0.10$, LEC versus FS utility is 0.602 versus 0.508 on CommonsenseQA, 0.298 versus 0.105 on TriviaQA, and 0.731 versus 0.648 on ScienceQA. Its corresponding pair-pooled SCERs are 10.13%, 10.31%, and 7.08%, versus 7.56%, 6.75%, and 7.01% for FS-Cert-U. The slight finite-test exceedances do not by themselves refute a population guarantee. These results do not establish universal utility superiority for FS-Cert-U: the contribution is certification of the policy set exposed to optimization, not a claim of dominating a retention-maximizing router with a richer search space.

Prior-routing comparison. For all 15 representative pair–dataset configurations, we use seeds 929000–929499 and exactly the original design/screen/certification/test indices. LEC reads only the certification examples for fitting, searches all effective calibration-score thresholds, and maximizes retention subject to $\sum_i (Z_i - \alpha S_i) \leq -1$. This is a reproduction of the published routing calibration rule, not a rerun of Wang's original model checkpoints or prompts. The shared count-grid implementation is checked against the original `fit_lec_constraint` implementation on the first split of every configuration at all six targets (90 comparisons). Tie-breaking follows the repository's retention, shifted-risk, and threshold-rank convention. LEC ignores the design/screen samples rather than reallocating them to calibration. Each repetition archives calibration inputs, threshold axes, and per-Primary-row feasible frontiers: these lossless calibration sufficient records reconstruct every grid count without storing billions of duplicate cells. The displayed comparison is not candidate-matched between LEC and FS-Cert-U; its search-space and guarantee differences must be retained when interpreting utility. Pointwise UCB baselines are not added as if they automatically provided simultaneous guarantees after grid search.

Reproduction notes. The existing local implementation is used, not claimed as a verified official upstream commit. No language-model outputs are regenerated. Source hashes and exact commands are archived in `FS-Cert-U/experiments/p0_p1_revision/README.md` and `source_versions.json`; the baseline entry point is `run_cached.py --task prior_baselines --workers 6`. All per-split results and reconstructable candidate counts remain unchanged.

Table 13: Matched representative-pair comparison at $\alpha = 0.10$, $\lambda = 0.25$: five pairs per dataset, 500 splits. LEC uses its full calibration-score grid and maximizes retention; other two-model methods share the frozen design family. SCER is pair-pooled; all other metrics include zero-output runs.

Method	SCER (%)	Cov. (%)	Guardian calls (%)	Utility	Nonempty (%)	Correct
CommonsenseQA						
Primary-FS	6.32	35.32	0.00	0.353	80.20	1172.6
Guardian-FS	6.47	59.74	100.00	0.347	100.00	1951.8
Naive	8.50	68.82	57.98	0.543	100.00	2205.0
Bonferroni	7.01	64.65	65.41	0.483	100.00	2102.8
Holm	7.06	64.86	65.18	0.486	100.00	2108.4
LEC-Routing	10.13	77.70	70.03	0.602	100.00	2444.5
FS-Cert-U	7.56	66.67	63.36	0.508	100.00	2155.5
TriviaQA						
Primary-FS	6.18	17.83	0.00	0.178	71.68	376.0
Guardian-FS	6.13	26.80	100.00	0.018	91.64	565.8
Naive	8.87	44.17	84.34	0.231	100.00	903.9
Bonferroni	5.78	26.37	87.13	0.046	89.92	558.8
Holm	5.80	26.44	87.05	0.047	89.92	560.1
LEC-Routing	10.31	50.87	84.45	0.298	100.00	1025.0
FS-Cert-U	6.75	31.59	84.48	0.105	94.76	661.5
ScienceQA						
Primary-FS	6.92	40.13	0.00	0.401	60.44	273.1
Guardian-FS	6.37	79.34	100.00	0.543	85.28	526.7
Naive	8.76	91.28	39.12	0.815	99.64	588.4
Bonferroni	6.92	67.67	61.75	0.522	69.44	448.0
Holm	7.00	67.86	61.36	0.525	69.44	448.8
LEC-Routing	7.08	93.52	81.65	0.731	99.44	614.6
FS-Cert-U	7.01	79.79	60.06	0.648	83.96	525.4

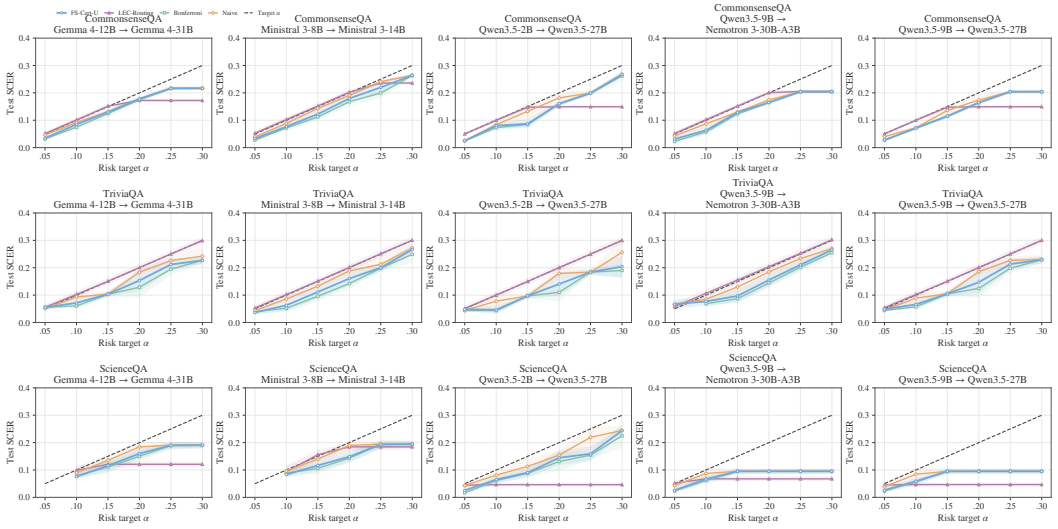


Figure 20: External routing comparison on three datasets and five representative pairs. Curves show mean and standard deviation of defined test SCER over 500 matched splits at each target. LEC uses its full calibration-score search space; the other methods share the frozen design family. The dashed line is the risk target, not a confidence bound. Table 13 reports coverage, Guardian calls, utility, nonempty outputs, and accepted correct counts.

E.9 Additional routing experiments: reproducibility and protocols

The revision bundle is `FS-Cert-U/experiments/p0_p1_revision/`. It contains separate prior-baseline, allocation, cost-tradeoff, non-monotonicity, and regret-stress outputs, including configurations, per-split records, numerical figure sources, and PDF/PNG figures. No new language-model inference or API calls are used.

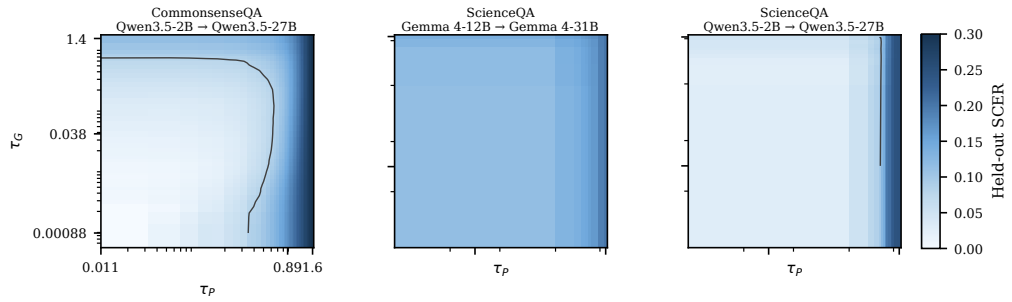


Figure 21: Additional archived held-out SCER surfaces for CommonsenseQA and ScienceQA. The contour marks SCER $= 0.10$; these comparative finite-pool displays complement the two TriviaQA witness panels in Figure 2 and do not establish global population non-monotonicity.

The five representative pairs are Gemma 4-12B→Gemma 4-31B, Ministral 3-8B→Ministral 3-14B, Qwen3.5-2B→Qwen3.5-27B, Qwen3.5-9B→Nemotron 3-30B-A3B, and Qwen3.5-9B→Qwen3.5-27B.

Cost sensitivity. All 33 main configurations use 500 matched splits and four fixed costs. Screen p-values are unchanged, but screen-utility tie-breaks and certification-utility maximization are recomputed for each cost. Every candidate's screen, certification, and test counts, p-values, and selected/certified IDs are retained. Test counts are used only after selection. The default-cost replay matches all original FS-Cert-U, Bonferroni, and Naive split records in answer, error, call counts and utility. The comparison includes empty-output runs. In the existing protocol the two-model all-abstain fallback has call rate one, Primary-FS has call rate zero, and Guardian-FS has call rate one; we preserve these conventions for numerical consistency.

Non-monotonicity witnesses. We inspect Qwen3.5-2B→Qwen3.5-27B, Qwen3.5-9B→Qwen3.5-27B, and Gemma 4-12B→Gemma 4-31B on all three main datasets at split seed 929000. Threshold grids are constructed exclusively from design scores. The heatmaps and witness search use the held-out test slice; they are exploratory displays, not independent tests of a preregistered population-monotonicity hypothesis. A witness must have coverage at least 0.10 and at least 50 answered examples at each of three ordered thresholds. We display the strongest Guardian-direction reversal and the strongest Primary-direction reversal from a different configuration, using the minimum of the two changes as strength. All searched surfaces, including weak or absent reversals, are retained. On TriviaQA 2B→27B, fixed $\tau_P = 0.666972$ and $\tau_G = (0.003754, 0.175260, 3.515754)$ give SCER $(0.383333, 0.176279, 0.375056)$ and answer counts $(300, 919, 2245)$. On TriviaQA 9B→27B, fixed $\tau_G = 0.059701$ and $\tau_P = (0.002759, 0.024529, 0.063163)$ give SCER $(0.032258, 0.027473, 0.035047)$ and answer counts $(310, 364, 428)$. The second reversal is much smaller; no population significance claim is made.

Near-tie controlled population. Three input types have weights $(0.50, 0.39, 0.11)$ and identical Primary and Guardian uncertainty vectors $(0.2, 0.5, 0.8)$. Primary error probabilities are $(0.02, 0.02, 0.8)$ and Guardian error probabilities are $(0.02, 0.02, 0.02)$; their errors are independent conditional on type. The two frozen policies are $(\tau_P, \tau_G) = (0.3, 0.9)$ and $(0.6, 0.3)$. Their coverages are 1 and 0.89, call rates 0.50 and 0.11, and utilities 0.875 and 0.8625. Both have SCER 0.02. For each m, NumPy SeedSequence $[2026100101, m]$ spawns independent screen/certification/test generators; sizes are 480, m, and $20m$. The original controlled-population candidate-count and selection functions are reused, with independent multinomial atom samples and 5,000 repetitions. Screening uses increasing screen p-value, then decreasing screen utility, then candidate ID. The family and population are fixed before the Monte Carlo run; no observed regret values are hard-coded or used to tune parameters. All atom counts and candidate-level decisions are archived. Empty certification, if present, would be recorded separately and excluded only from within-certified-set regret, not from nonempty/safe-output rates. This experiment tests utility selection, not an unsafe-policy boundary; the original FWER study remains separate.

E.10 Additional utility and paired-gain curves

Figure 22 reports the coverage curves corresponding to the four methods in main-text Figure 4. Figure 23 additionally retains the archived split-variability bands for FS-Cert-U, Bonferroni, and Naive.

E.11 Accepted correct answers for every model pair

Tables 14–17 report all 44 model pairs, five methods, and six risk targets. Each entry averages the number of accepted correct test answers across 500 splits, including zero-output runs. Test sizes are 3,500 for CommonsenseQA, 2,246 for TriviaQA, 706 for ScienceQA, and 181 for MM-Vet. In the last dataset, correctness refers to

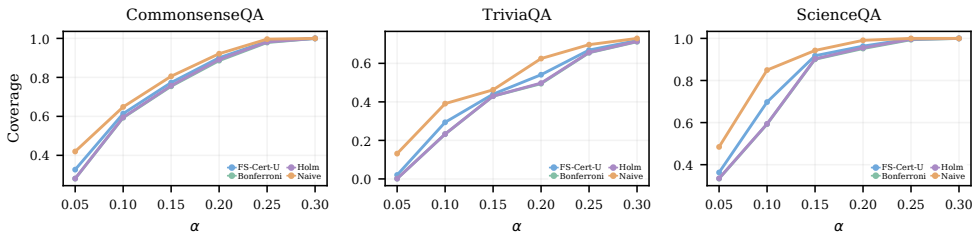


Figure 22: Coverage across all six risk targets at $\lambda = 0.25$ for FS-Cert-U, Bonferroni, Holm, and Naive. Means average equally over the same 11 pairs per dataset and 500 matched splits as main-text Figure 4.

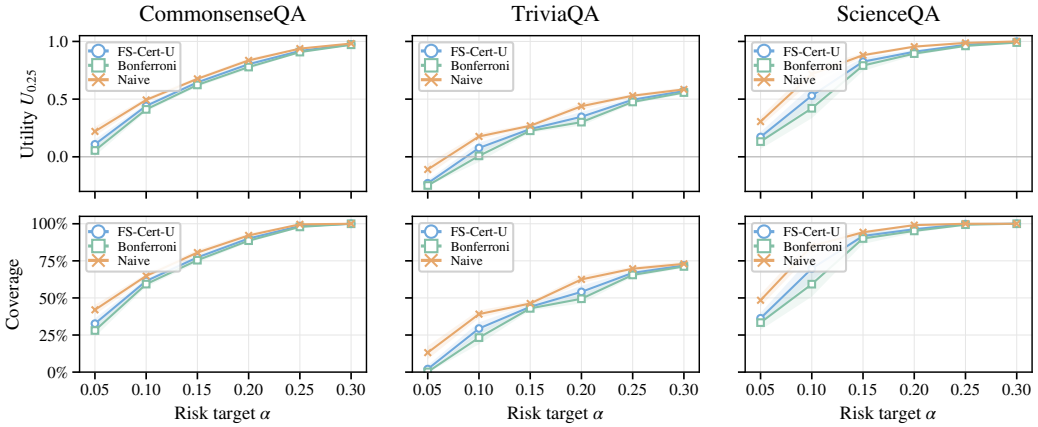


Figure 23: Utility and coverage across risk targets. Each curve averages the 11 model pairs within a split, then summarizes 500 splits; shading denotes one standard deviation. Utility charges every Guardian call, including calls followed by abstention.

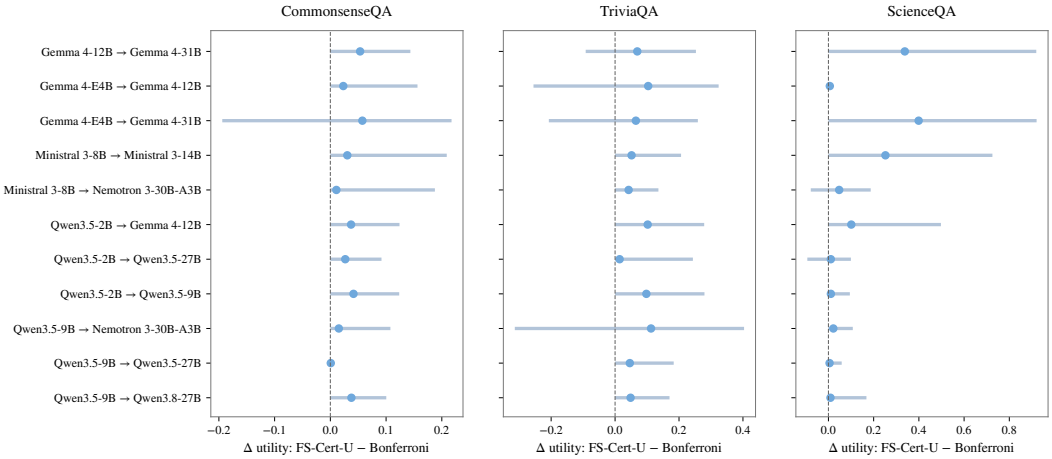


Figure 24: Paired utility differences at $\alpha = 0.10$ for all 33 main model pairs. Points denote the mean FS-Cert-U minus Bonferroni difference over 500 matched splits; segments denote the empirical 2.5th–97.5th percentile range. The vertical dashed line marks zero.

the sentence-similarity criterion. The accompanying experiment artifacts retain the individual split records and standard deviations.

Table 14: Accepted correct test answers: CommonsenseQA. Mean over 500 splits; columns are risk targets α.

Method	0.05	0.10	0.15	0.20	0.25	0.30
Gemma 4-12B → Gemma 4-31B						
Primary-FS	837.4	1651.7	2285.7	2296.4	2742.0	2742.0
Guardian-FS	753.3	2309.9	2406.0	2899.3	2899.3	2899.3
Bonferroni	1048.0	2213.4	2528.5	2869.6	2743.7	2742.0
Naive	1727.3	2448.8	2600.0	2876.7	2742.0	2742.0
FS-Cert-U	1294.4	2375.1	2585.5	2875.9	2742.0	2742.0
Gemma 4-E4B → Gemma 4-12B						
Primary-FS	355.1	997.8	1584.1	2164.2	2173.7	2573.7
Guardian-FS	837.4	1651.7	2285.7	2296.4	2742.0	2742.0
Bonferroni	705.0	1708.5	2247.8	2353.4	2713.1	2576.1
Naive	1181.0	1856.4	2370.5	2491.4	2697.5	2573.7
FS-Cert-U	907.8	1723.3	2335.3	2441.1	2701.4	2573.7
Gemma 4-E4B → Gemma 4-31B						
Primary-FS	355.1	997.8	1584.1	2164.2	2173.7	2573.7
Guardian-FS	753.3	2309.9	2406.0	2899.3	2899.3	2899.3
Bonferroni	322.4	2110.2	2465.4	2789.0	2778.6	2577.9
Naive	1636.7	2412.6	2463.2	2851.1	2775.1	2573.7
FS-Cert-U	715.3	2266.0	2465.4	2854.9	2778.6	2573.7
Ministral 3-8B → Ministral 3-14B						
Primary-FS	481.6	880.1	1571.2	2080.2	2154.1	2577.0
Guardian-FS	850.0	1621.3	1898.0	2261.1	2534.8	2676.1
Bonferroni	779.1	1591.6	1924.1	2280.8	2471.0	2577.8
Naive	1048.0	1668.4	2262.7	2409.1	2655.7	2577.0
FS-Cert-U	902.5	1654.5	2044.2	2305.9	2569.6	2577.0
Ministral 3-8B → Nemotron 3-30B-A3B						
Primary-FS	481.6	880.1	1571.2	2080.2	2154.1	2577.0
Guardian-FS	311.4	898.2	1580.1	2092.4	2156.9	2562.5
Bonferroni	149.2	1152.8	1880.2	2184.6	2467.5	2577.5
Naive	1032.5	1513.9	1975.0	2340.1	2567.7	2577.0
FS-Cert-U	424.9	1181.5	1962.5	2257.8	2470.7	2577.0
Qwen3.5-2B → Gemma 4-12B						
Primary-FS	0.0	7.4	777.2	828.3	1402.2	1630.3
Guardian-FS	837.4	1651.7	2285.7	2296.4	2742.0	2742.0
Bonferroni	689.8	1665.4	2202.7	2320.6	2721.3	2638.6
Naive	944.8	1753.0	2317.3	2335.4	2657.7	2497.7
FS-Cert-U	837.4	1699.8	2296.6	2325.4	2717.9	2606.9
Qwen3.5-2B → Qwen3.5-27B						
Primary-FS	0.0	7.4	777.2	828.3	1402.2	1630.3
Guardian-FS	1706.1	2464.8	2477.7	2977.2	2977.2	2977.2
Bonferroni	1706.1	2466.9	2468.2	2942.4	2803.7	2583.7
Naive	1713.4	2468.3	2756.5	2864.8	2801.0	2563.4
FS-Cert-U	1706.1	2467.9	2475.0	2938.0	2803.7	2563.9
Qwen3.5-2B → Qwen3.5-9B						
Primary-FS	0.0	7.4	777.2	828.3	1402.2	1630.3
Guardian-FS	879.2	1662.0	2327.7	2328.5	2784.7	2784.7
Bonferroni	862.0	1704.4	2335.5	2342.3	2755.5	2655.5
Naive	1225.4	1774.0	2341.9	2445.7	2668.0	2498.7
FS-Cert-U	879.2	1742.4	2340.9	2343.7	2718.6	2610.0
Qwen3.5-9B → Nemotron 3-30B-A3B						
Primary-FS	879.2	1662.0	2327.7	2328.5	2784.7	2784.7
Guardian-FS	311.4	898.2	1580.1	2092.4	2156.9	2562.5
Bonferroni	949.1	1740.6	2385.7	2568.8	2784.7	2784.7
Naive	1386.1	1938.7	2422.1	2617.4	2784.7	2784.7
FS-Cert-U	1083.1	1778.9	2415.9	2569.1	2784.7	2784.7
Qwen3.5-9B → Qwen3.5-27B						
Primary-FS	879.2	1662.0	2327.7	2328.5	2784.7	2784.7
Guardian-FS	1706.1	2464.8	2477.7	2977.2	2977.2	2977.2
Bonferroni	1735.3	2501.2	2628.0	2925.1	2784.7	2784.7
Naive	1845.4	2500.8	2823.0	2891.5	2784.7	2784.7
FS-Cert-U	1742.1	2501.2	2632.7	2924.9	2784.7	2784.7

Continued on next page

Continued: CommonsenseQA

Method	0.05	0.10	0.15	0.20	0.25	0.30
Qwen3.5-9B → Qwen3.8-27B						
Primary-FS	879.2	1662.0	2327.7	2328.5	2784.7	2784.7
Guardian-FS	1657.8	2419.9	2425.7	2921.6	2921.6	2921.6
Bonferroni	1556.3	2411.2	2602.2	2896.4	2784.7	2784.7
Naive	1762.4	2474.9	2612.9	2868.7	2784.7	2784.7
FS-Cert-U	1698.5	2467.3	2612.5	2895.5	2784.7	2784.7

Table 15: Accepted correct test answers: TriviaQA. Mean over 500 splits; columns are risk targets α.

Method	0.05	0.10	0.15	0.20	0.25	0.30
Gemma 4-12B → Gemma 4-31B						
Primary-FS	0.0	455.3	525.2	583.9	908.3	908.3
Guardian-FS	21.5	660.5	1021.2	1272.4	1382.8	1383.1
Bonferroni	1.1	617.1	1055.9	1144.5	1381.4	1377.2
Naive	325.7	999.2	1056.9	1380.0	1377.5	1391.0
FS-Cert-U	15.4	743.1	1056.7	1242.2	1379.5	1377.2
Gemma 4-E4B → Gemma 4-12B						
Primary-FS	0.0	284.7	521.7	523.0	722.7	871.4
Guardian-FS	0.0	455.3	525.2	583.9	908.3	908.3
Bonferroni	0.0	277.2	705.4	718.9	957.7	1048.5
Naive	37.3	677.9	709.5	911.6	997.6	1077.3
FS-Cert-U	0.0	451.2	707.9	764.3	953.8	1070.4
Gemma 4-E4B → Gemma 4-31B						
Primary-FS	0.0	284.7	521.7	523.0	722.7	871.4
Guardian-FS	21.5	660.5	1021.2	1272.4	1382.8	1383.1
Bonferroni	1.1	592.3	1089.3	1167.9	1389.8	1365.6
Naive	314.5	1006.4	1093.7	1387.9	1373.5	1388.5
FS-Cert-U	18.4	703.5	1093.1	1286.7	1385.8	1364.8
Ministral 3-8B → Ministral 3-14B						
Primary-FS	23.5	538.8	768.7	976.7	978.3	1218.3
Guardian-FS	253.1	730.8	1026.1	1203.5	1375.7	1375.7
Bonferroni	27.7	705.9	1053.4	1186.5	1367.2	1329.6
Naive	599.4	1026.9	1137.8	1359.4	1350.1	1343.5
FS-Cert-U	220.4	806.0	1088.4	1266.4	1365.2	1335.2
Ministral 3-8B → Nemotron 3-30B-A3B						
Primary-FS	23.5	538.8	768.7	976.7	978.3	1218.3
Guardian-FS	0.0	299.3	602.4	963.9	988.4	1266.8
Bonferroni	1.0	638.2	815.5	1153.5	1191.7	1301.9
Naive	330.2	762.9	1072.7	1185.1	1294.8	1334.0
FS-Cert-U	21.5	722.0	894.3	1181.0	1210.6	1307.6
Qwen3.5-2B → Gemma 4-12B						
Primary-FS	0.0	0.0	0.0	0.0	0.0	0.0
Guardian-FS	0.0	455.3	525.2	583.9	908.3	908.3
Bonferroni	0.0	241.4	525.2	532.9	887.8	908.3
Naive	32.6	525.2	525.2	789.3	908.3	906.8
FS-Cert-U	0.0	455.3	525.2	583.9	908.3	908.3
Qwen3.5-2B → Qwen3.5-27B						
Primary-FS	0.0	0.0	0.0	0.0	0.0	0.0
Guardian-FS	54.4	569.3	1015.2	1196.5	1373.2	1373.2
Bonferroni	1.0	542.2	1013.3	1071.6	1373.2	1364.3
Naive	447.0	841.7	1015.2	1354.7	1373.2	1266.3
FS-Cert-U	54.4	569.3	1015.2	1196.5	1373.2	1343.7
Qwen3.5-2B → Qwen3.5-9B						
Primary-FS	0.0	0.0	0.0	0.0	0.0	0.0
Guardian-FS	1.9	442.9	524.6	621.7	908.3	911.6
Bonferroni	0.0	238.4	524.6	544.1	895.7	911.6
Naive	36.9	521.1	525.2	821.4	911.6	908.5
FS-Cert-U	1.9	442.9	524.6	621.7	908.3	911.3

Continued on next page

Continued: TriviaQA

Method	0.05	0.10	0.15	0.20	0.25	0.30
Qwen3.5-9B → Nemotron 3-30B-A3B						
Primary-FS	1.9	442.9	524.6	621.7	908.3	911.6
Guardian-FS	0.0	299.3	602.4	963.9	988.4	1266.8
Bonferroni	0.0	267.3	754.7	1024.0	1103.2	1282.9
Naive	36.9	733.4	965.7	1037.1	1236.5	1278.6
FS-Cert-U	1.9	464.6	804.7	1021.4	1140.7	1281.2
Qwen3.5-9B → Qwen3.5-27B						
Primary-FS	1.9	442.9	524.6	621.7	908.3	911.6
Guardian-FS	54.4	569.3	1015.2	1196.5	1373.2	1373.2
Bonferroni	1.0	661.5	1039.1	1094.5	1359.4	1324.9
Naive	455.4	918.2	1040.8	1354.1	1327.8	1322.5
FS-Cert-U	50.4	724.6	1040.9	1173.9	1342.7	1324.9
Qwen3.5-9B → Qwen3.8-27B						
Primary-FS	1.9	442.9	524.6	621.7	908.3	911.6
Guardian-FS	106.1	553.4	1007.7	1008.7	1225.5	1305.9
Bonferroni	1.9	621.7	1023.5	1035.2	1169.3	1282.9
Naive	478.1	823.8	1032.0	1082.8	1295.1	1272.2
FS-Cert-U	101.8	682.1	1031.1	1045.7	1238.0	1275.9

Table 16: Accepted correct test answers: ScienceQA. Mean over 500 splits; columns are risk targets α.

Method	0.05	0.10	0.15	0.20	0.25	0.30
Gemma 4-12B → Gemma 4-31B						
Primary-FS	0.0	5.8	479.6	497.9	571.2	571.6
Guardian-FS	0.0	462.6	613.3	620.1	620.1	620.1
Bonferroni	0.0	226.3	545.0	590.2	572.7	571.6
Naive	0.0	544.5	582.4	575.7	571.6	571.6
FS-Cert-U	0.0	446.2	559.5	591.3	571.7	571.6
Gemma 4-E4B → Gemma 4-12B						
Primary-FS	0.0	0.0	154.9	435.1	452.2	510.4
Guardian-FS	0.0	5.8	479.6	497.9	571.2	571.6
Bonferroni	0.0	0.0	390.0	481.1	531.7	540.1
Naive	0.0	214.6	461.8	522.4	543.9	523.9
FS-Cert-U	0.0	3.8	421.1	497.0	546.5	528.8
Gemma 4-E4B → Gemma 4-31B						
Primary-FS	0.0	0.0	154.9	435.1	452.2	510.4
Guardian-FS	0.0	462.6	613.3	620.1	620.1	620.1
Bonferroni	0.0	201.5	599.1	563.4	570.7	548.3
Naive	0.0	591.9	567.3	573.7	554.6	523.9
FS-Cert-U	0.0	462.6	598.1	569.7	566.7	531.2
Ministral 3-8B → Ministral 3-14B						
Primary-FS	0.0	6.2	461.2	490.9	568.3	568.9
Guardian-FS	0.0	221.2	491.3	497.1	569.2	569.8
Bonferroni	0.0	40.2	473.1	533.8	568.3	568.9
Naive	0.0	458.1	529.4	567.0	568.9	568.9
FS-Cert-U	0.0	202.8	490.0	538.8	568.5	568.9
Ministral 3-8B → Nemotron 3-30B-A3B						
Primary-FS	0.0	6.2	461.2	490.9	568.3	568.9
Guardian-FS	43.6	605.0	648.8	648.8	648.8	648.8
Bonferroni	2.2	571.2	609.7	617.0	573.9	568.9
Naive	452.8	608.5	617.3	581.9	568.9	568.9
FS-Cert-U	43.6	592.9	615.5	614.6	569.1	568.9
Qwen3.5-2B → Gemma 4-12B						
Primary-FS	14.4	268.9	333.4	387.9	446.9	532.5
Guardian-FS	0.0	5.8	479.6	497.9	571.2	571.6
Bonferroni	1.1	212.4	488.2	526.3	569.1	538.7
Naive	194.9	422.5	520.6	571.7	541.2	533.4
FS-Cert-U	14.4	266.3	502.7	539.2	561.2	533.7

Continued on next page

Continued: ScienceQA

Method	0.05	0.10	0.15	0.20	0.25	0.30
Qwen3.5-2B → Qwen3.5-27B						
Primary-FS	14.4	268.9	333.4	387.9	446.9	532.5
Guardian-FS	551.4	672.5	672.5	672.5	672.5	672.5
Bonferroni	547.0	662.7	643.9	613.3	597.0	547.3
Naive	629.2	648.7	626.2	596.9	551.3	533.4
FS-Cert-U	549.1	659.0	642.3	604.0	593.6	534.0
Qwen3.5-2B → Qwen3.5-9B						
Primary-FS	14.4	268.9	333.4	387.9	446.9	532.5
Guardian-FS	522.0	542.3	638.8	638.8	638.8	638.8
Bonferroni	430.7	538.5	629.4	614.2	583.2	544.2
Naive	533.1	603.2	618.1	584.4	547.2	533.4
FS-Cert-U	498.1	541.0	624.3	602.7	579.9	533.9
Qwen3.5-9B → Nemotron 3-30B-A3B						
Primary-FS	522.0	542.3	638.8	638.8	638.8	638.8
Guardian-FS	43.6	605.0	648.8	648.8	648.8	648.8
Bonferroni	467.1	645.4	639.0	638.8	638.8	638.8
Naive	595.0	644.3	638.8	638.8	638.8	638.8
FS-Cert-U	517.8	654.6	638.8	638.8	638.8	638.8
Qwen3.5-9B → Qwen3.5-27B						
Primary-FS	522.0	542.3	638.8	638.8	638.8	638.8
Guardian-FS	551.4	672.5	672.5	672.5	672.5	672.5
Bonferroni	576.1	665.5	639.1	638.8	638.8	638.8
Naive	617.9	646.5	638.8	638.8	638.8	638.8
FS-Cert-U	582.8	664.3	638.8	638.8	638.8	638.8
Qwen3.5-9B → Qwen3.8-27B						
Primary-FS	522.0	542.3	638.8	638.8	638.8	638.8
Guardian-FS	479.1	520.3	623.4	627.6	627.6	627.6
Bonferroni	508.3	581.7	638.6	638.8	638.8	638.8
Naive	579.1	629.6	638.8	638.8	638.8	638.8
FS-Cert-U	534.3	584.4	638.9	638.8	638.8	638.8

Table 17: Accepted correct test answers: MM-Vet v2. Mean over 500 splits; columns are risk targets α.

Method	0.05	0.10	0.15	0.20	0.25	0.30
Gemma 4-12B → Gemma 4-31B						
Primary-FS	0.0	0.7	8.3	29.0	62.6	107.6
Guardian-FS	0.0	2.0	19.6	65.3	125.3	146.7
Bonferroni	0.0	0.0	3.5	29.8	84.2	135.4
Naive	18.7	60.6	97.3	136.4	139.1	137.1
FS-Cert-U	0.0	1.3	15.4	52.6	105.1	135.4
Gemma 4-E4B → Gemma 4-12B						
Primary-FS	0.0	0.1	3.2	12.0	30.2	52.0
Guardian-FS	0.0	0.7	8.3	29.0	62.6	107.6
Bonferroni	0.0	0.0	0.5	9.4	36.4	77.4
Naive	9.7	38.4	66.8	91.2	120.8	128.2
FS-Cert-U	0.0	0.5	6.5	26.6	57.0	91.0
Gemma 4-E4B → Gemma 4-31B						
Primary-FS	0.0	0.1	3.2	12.0	30.2	52.0
Guardian-FS	0.0	2.0	19.6	65.3	125.3	146.7
Bonferroni	0.0	0.0	2.9	26.1	80.7	135.7
Naive	15.6	56.7	94.7	136.6	137.8	130.1
FS-Cert-U	0.0	1.5	13.5	52.2	104.5	134.9
Ministral 3-8B → Gemma 4-31B						
Primary-FS	0.0	0.0	0.4	3.7	15.1	35.6
Guardian-FS	0.0	2.0	19.6	65.3	125.3	146.7
Bonferroni	0.0	0.0	2.5	24.0	78.9	135.5
Naive	13.7	53.2	94.2	136.8	138.7	131.8
FS-Cert-U	0.0	1.4	16.7	52.9	104.1	135.4

Continued on next page

Continued: MM-Vet v2

Method	0.05	0.10	0.15	0.20	0.25	0.30
Ministral 3-8B → Ministral 3-14B						
Primary-FS	0.0	0.0	0.4	3.7	15.1	35.6
Guardian-FS	0.0	0.2	0.8	5.3	18.6	50.6
Bonferroni	0.0	0.0	0.1	0.7	6.1	30.4
Naive	1.7	11.4	37.2	66.2	96.7	121.5
FS-Cert-U	0.0	0.1	0.8	7.2	23.7	51.7
Qwen3.5-2B → Gemma 4-12B						
Primary-FS	0.0	0.0	0.1	2.3	16.6	59.4
Guardian-FS	0.0	0.7	8.3	29.0	62.6	107.6
Bonferroni	0.0	0.0	0.1	7.4	32.8	79.3
Naive	7.7	33.1	66.4	95.3	123.6	131.8
FS-Cert-U	0.0	0.4	6.6	24.2	51.4	89.6
Qwen3.5-2B → Qwen3.5-27B						
Primary-FS	0.0	0.0	0.1	2.3	16.6	59.4
Guardian-FS	0.0	0.0	0.9	9.9	48.8	106.7
Bonferroni	0.0	0.0	0.0	1.4	16.1	67.4
Naive	1.4	13.7	50.7	94.6	124.2	132.0
FS-Cert-U	0.0	0.0	0.4	6.1	32.7	82.3
Qwen3.5-2B → Qwen3.5-9B						
Primary-FS	0.0	0.0	0.1	2.3	16.6	59.4
Guardian-FS	0.0	0.1	1.1	14.5	64.8	113.1
Bonferroni	0.0	0.0	0.0	2.1	25.0	82.5
Naive	1.2	15.3	57.9	101.8	124.5	131.9
FS-Cert-U	0.0	0.1	0.9	9.1	45.8	93.2
Qwen3.5-9B → Gemma 4-31B						
Primary-FS	0.0	0.1	1.1	14.5	64.8	113.1
Guardian-FS	0.0	2.0	19.6	65.3	125.3	146.7
Bonferroni	0.0	0.0	2.6	26.1	87.0	136.1
Naive	14.4	56.3	101.9	137.5	139.1	137.5
FS-Cert-U	0.0	1.3	13.6	48.2	104.6	135.6
Qwen3.5-9B → Qwen3.5-27B						
Primary-FS	0.0	0.1	1.1	14.5	64.8	113.1
Guardian-FS	0.0	0.0	0.9	9.9	48.8	106.7
Bonferroni	0.0	0.0	0.0	3.0	32.4	94.7
Naive	2.3	20.3	67.6	111.7	133.5	137.4
FS-Cert-U	0.0	0.1	1.0	9.3	55.3	110.8
Qwen3.5-9B → Qwen3.8-27B						
Primary-FS	0.0	0.1	1.1	14.5	64.8	113.1
Guardian-FS	0.0	1.7	12.5	42.1	87.2	128.2
Bonferroni	0.0	0.0	1.3	12.9	58.1	111.0
Naive	10.6	45.0	84.8	118.1	135.2	137.4
FS-Cert-U	0.0	1.0	9.1	31.9	79.0	122.3

E.12 Expanded main comparisons and secondary diagnostics

Full evaluation protocol.

Benchmarks and models. Our main evaluation uses CommonsenseQA (Talmor et al., 2019), TriviaQA (Joshi et al., 2017), and ScienceQA (Lu et al., 2022), containing 10,000, 6,416, and 2,017 distinct questions, respectively. CommonsenseQA and ScienceQA use multiple-choice correctness; TriviaQA uses normalized exact match against reference answers and aliases. We evaluate 11 directed model pairs per dataset, covering ten models from the Qwen3.5 (Qwen Team, 2026b), Qwen3.8 (Qwen Team, 2026a), Gemma 4 (Gemma Team, 2026), Ministral 3 (Liu et al., 2026), and Nemotron 3 Nano Omni (NVIDIA, 2026) families. The pairs include both within-family and cross-family routing. An additional evaluation on MM-Vet v2 (Yu et al., 2024) uses 516 questions and 11 pairs of nine models, with sentence-similarity correctness labels. Model pairs and generation settings are listed in Appendix E.1.

Calibration protocol. Each main benchmark comparison uses 500 random question-level splits, shared by its compared methods. The design, screen, certification, and test fractions are 10%, 20%, 35%, and 35%. The default family uses the design-score quartiles and infinite endpoints, yielding $K = 7$–21 distinct routing

candidates. Screen ordering uses increasing screen p-value, then decreasing screen utility. We set $\delta = 0.05$ and $\lambda = 0.25$, and evaluate $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. Model outputs and uncertainty scores are held fixed across calibration splits; the separately described sensitivity and controlled-population studies use their own repetition counts.

Comparison classes. Our primary comparisons isolate certification and selection by holding the frozen two-model candidate family and matched splits fixed across **FS-Cert-U**, **Bonferroni**, **Holm**, and **Naive**. Bonferroni applies the same exact tests at level δ/K; Holm applies step-down testing to the same p-values. Both maximize certification utility within their accepted sets. Naive maximizes empirical utility subject to empirical risk, without multiplicity correction, and does not provide the family-wise certificate.

Primary-FS and **Guardian-FS** are single-model references, using their marginal threshold families on matched splits. They assess joint routing relative to separately certified models; Primary-FS makes no Guardian calls, while Guardian-FS invokes it on every query.

An additional comparison with **LEC-Routing** is reported in Appendix E.8. Because LEC uses a richer calibration-score search space and targets a different risk statement, we treat it as an external routing reference rather than a controlled certification ablation.

Metrics. We report test SCER, coverage, Guardian call rate, utility U_λ (with the default $\lambda = 0.25$), and the number of accepted correct answers. Routing allocation is the fraction of all test questions answered by each model. Curves show means and standard deviations over splits; dataset summaries weight model pairs equally. Coverage and utility include every split, while SCER is defined only for nonempty answer sets. We also report the frequency of test SCER exceeding α among nonempty splits as a measure of variation across calibrations. Exact aggregation formulas are given in Appendix E.2.

Detailed landscape observations. Figure 2 evaluates a 41×41 design-quantile grid (ties removed) on a fixed held-out TriviaQA split. Configurations were chosen descriptively from nine prelisted pair–dataset combinations, without feeding these test results into certification. At fixed $\tau_P = 0.667$, increasing the Guardian threshold gives SCER $0.383 \to 0.176 \to 0.375$ for Qwen3.5-2B→Qwen3.5-27B, with coverage $0.134 \to 0.409 \to 1.000$. For Qwen3.5-9B→Qwen3.5-27B at fixed $\tau_G = 0.0597$, increasing the Primary threshold gives $0.0323 \to 0.0275 \to 0.0350$, based on 310, 364, and 428 answered examples. The Primary-direction reversal is small and may be sensitive to sampling; these are empirical finite-pool witnesses, not a claim that the population surface is everywhere non-monotonic. The Guardian-threshold example is pronounced, whereas the Primary-threshold reversal is smaller and may be sampling-sensitive. These observations do not establish global population non-monotonicity. They show why monotonicity should not be assumed from the routing architecture alone, motivating direct policy certification. Numerical thresholds, action changes, and all nine searched surfaces are archived.

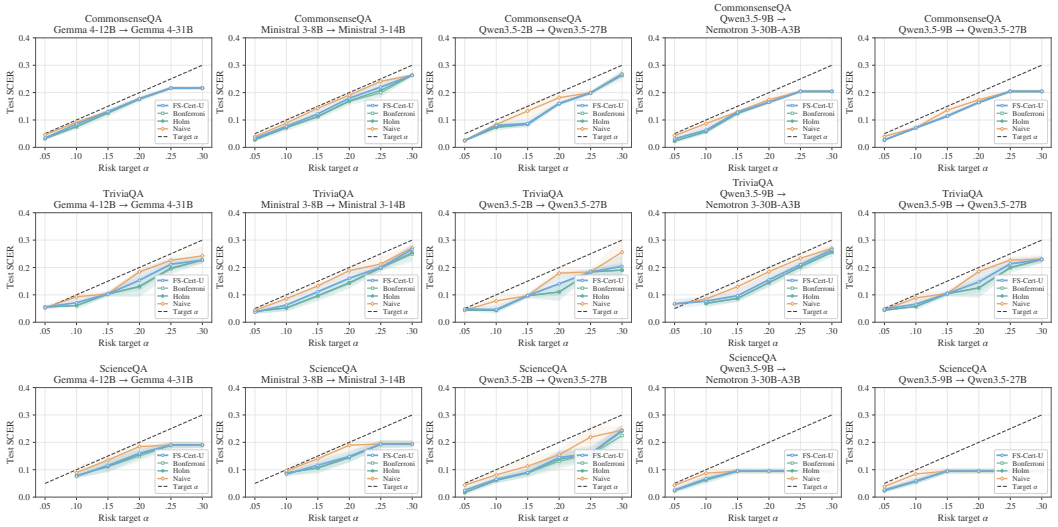


Figure 25: Test SCER for three datasets and five routing pairs (mean ± standard deviation over nonempty test splits; 500 matched splits). FS-Cert-U, Bonferroni, Holm, and Naive use the same frozen design family and matched splits. The dashed diagonal is the risk target, not a confidence bound.

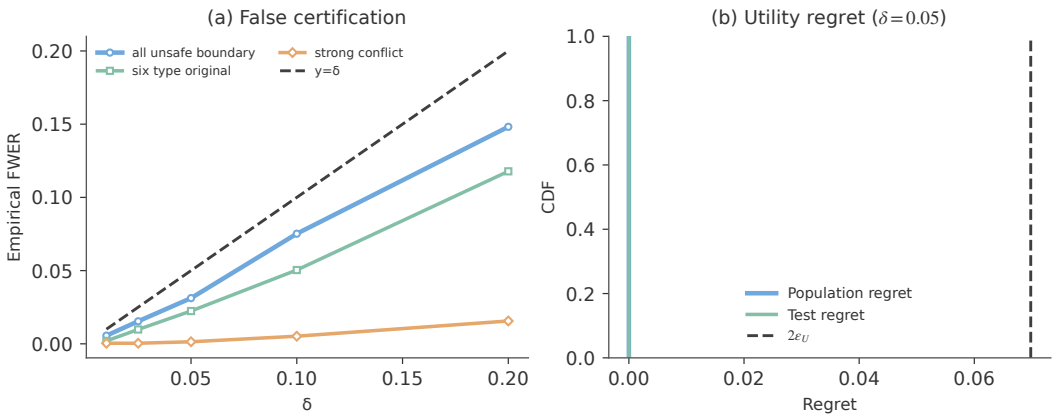


Figure 26: Theorem-aligned validation. (a) FS-Cert-U empirical false-certification probability versus δ at $m = 4480$; the dashed diagonal is the nominal family-wise budget. (b) CDFs of population and independent-test regret for FS-Cert-U at $m = 4480$ and $\delta = 0.05$, relative to the best policy inside the certified set; the vertical line is the corresponding $2\varepsilon_U$. The regret curves pool the three controlled populations. The underlying replay also evaluates Holm and Bonferroni.

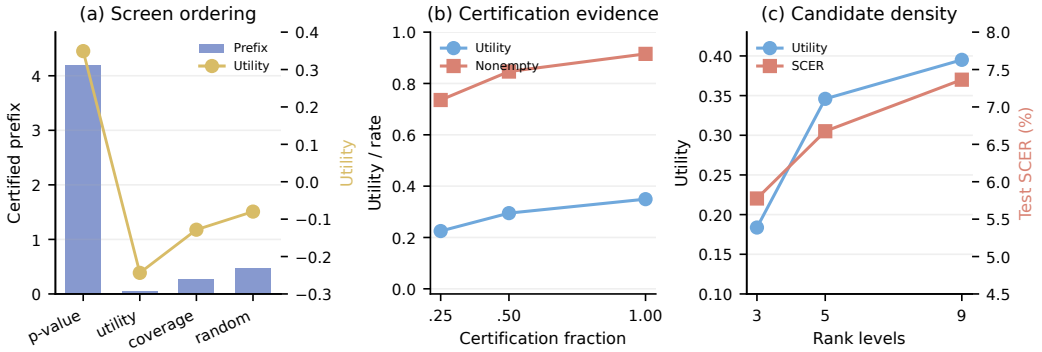


Figure 27: **What determines certification power?** (a) Screen ordering controls the certified prefix; points show the resulting utility. (b) More certification evidence increases utility and the chance of a nonempty certified set. (c) A richer candidate family increases attainable utility while the test SCER remains near the target. Values are equal-weight means over the 33 main configurations at $\alpha = 0.10$; (a–b) use 500 matched splits and (c) uses the 50-split sensitivity replay.

Figure 25 compares system-level risk across the six targets. FS-Cert-U generally operates closer to the target than Bonferroni, which imposes a smaller testing level on every candidate. Table 4 summarizes the results at $\alpha = 0.10$: FS-Cert-U obtains SCERs of 7.41%, 6.86%, and 6.80% on CommonsenseQA, TriviaQA, and ScienceQA, respectively. Bonferroni reduces these values further, to 6.73%, 6.07%, and 6.08%, with lower coverage on each dataset. Lower empirical SCER alone is not superiority: Bonferroni operates farther below the target, consistent with a more conservative rule. The relevant comparison is how much useful behavior is retained while targeting the same family-wise certification event.

Average risk alone does not capture variation across splits. Naive attains greater coverage, but its test SCER exceeds the target in 12.76%, 15.35%, and 33.14% of nonempty splits on the three datasets. FS-Cert-U reduces these frequencies to 1.76%, 2.50%, and 2.79%. The contrast is most pronounced on ScienceQA, which has the smallest certification split among the three main benchmarks. Bonferroni has still fewer exceedances, consistent with its more conservative operating point. Appendix E.3 reports exceedance and zero-output rates for every target.

Coverage–cost trade-off. The difference in certification power translates into higher utility. At $\alpha = 0.10$, FS-Cert-U improves mean utility over Bonferroni by 0.0304 on CommonsenseQA, 0.0685 on TriviaQA, and 0.1091 on ScienceQA. Coverage increases by 2.09, 6.15, and 10.48 percentage points, respectively, while Guardian call rates decrease by 3.78, 2.78, and 1.75 points. The paired mean utility improvement is positive for all 33 model pairs. Appendix Figure 23 shows that the dataset-level utility advantage extends across all six targets, with smaller differences as the risk constraint becomes less restrictive.

Certification also affects how often a usable policy is returned. At $\alpha = 0.10$, FS-Cert-U produces nonempty test outputs in 100%, 93.00%, and 81.58% of splits, compared with 100%, 81.36%, and 69.35% for Bonferroni. These are nonempty test-output frequencies, not direct measurements of certified-set nonemptiness. Table 4 also reports greater coverage and more accepted correct answers for FS-Cert-U than Bonferroni and Holm on each dataset.

The improvement is not confined to one correction baseline: at $\alpha = 0.10$, mean utility for FS-Cert-U, Holm, and Bonferroni is respectively $(0.441, 0.414, 0.411)$ on CommonsenseQA, $(0.077, 0.009, 0.008)$ on TriviaQA, and $(0.530, 0.422, 0.421)$ on ScienceQA. Across all 33 main model-pair configurations, the paired FS–Bonferroni utility difference is positive. This supports a certification-power claim within the matched frozen-family setting, not universal routing dominance. Representative routing allocation and correct-answer recovery are analyzed below; complete all-pair results appear in Appendix E.3.

E.12.1 Set-level safety event

The benchmark replay measures finite-test behavior, whereas the theory controls a more specific event: an unsafe policy entering the certified set. We therefore run a controlled-population replay with 27 frozen candidates, three six-type error scenarios (original, conflict, and all-unsafe boundary), a two-million-example oracle pool, and 5,000 independent splits per setting. A candidate is oracle-unsafe when its large-pool SCER estimate is at least $\alpha = 0.10$.

Panel (a) estimates $R^{-1} \sum_{r=1}^{R} \mathbf{1}\{\exists \theta_k \in \mathcal{A}_{\text{cert}}^{(r)} : q_k^{\text{oracle}} \geq \alpha\}$, where q_k^{oracle} is the large-pool estimate of the population SCER for frozen candidate k. The empirical rates remain below the nominal δ for every displayed setting, directly targeting the event controlled by Theorem 1, with the stated oracle-pool approximation. This is not the event that every finite test split has SCER below α. Panel (b) compares the selected policy with the best accepted policy using an independent test set. Both population and test regret are exactly zero in the displayed replay and hence below $2\varepsilon_U$; this setting evaluates the within-certified-set comparison but does not probe the bound's sharpness. The oracle-pool approximation has maximum SCER error 7.2×10^{-4} (Appendix E.6).

Power mechanisms. We vary the certification fraction, candidate density, screen ordering, and Guardian cost using matched splits. Figure 27 makes the mechanism visible: the p-value-first screen order yields a certified prefix of 4.19 candidates, compared with 0.04 for utility-first ordering; increasing certification data from one quarter to the full split raises utility from 0.225 to 0.349; and increasing the threshold grid from three to nine levels raises utility from 0.184 to 0.395 while keeping test SCER on the same scale. These changes affect power and attainable utility, not the validity theorem.

The gain over Bonferroni is systematic across model pairs: all 33 paired mean differences are positive at $\alpha = 0.10$ (Appendix Figure 24). This is a matched frozen-family power comparison, not universal routing dominance.

MM-Vet v2 stress test. An additional open-ended visual-answer study uses proxy sentence-similarity labels and only 181 certification questions per split. Its strict-target low-coverage regime, including the $\alpha = 0.25$ comparison where FS-Cert-U reaches 46.89% coverage and 0.2719 utility, is reported in Appendix E.7; it is intended as a stress test of certification sample size rather than as the primary theorem evidence.

E.12.2 Routing allocation

Figures 28–29 inspect Qwen3.5-2B→Qwen3.5-9B on 500 matched splits. At $\alpha = 0.10$, the Primary/Guardian answered fractions for FS-Cert-U are $19.59\%/34.29\%$ on CommonsenseQA, $0\%/21.14\%$ on TriviaQA, and $42.18\%/38.87\%$ on ScienceQA. Guardian call rates are 80.41%, 100%, and 57.82%, respectively: invocation includes escalation followed by abstention, and must not be equated with Guardian acceptance. Correct-answer recovery is 58.63%, 35.69%, and 81.18%. FS-Cert-U does not constrain the selected policy to use both stages non-trivially. On CommonsenseQA and ScienceQA, selected policies use both answering stages substantially. On this TriviaQA configuration, the certification-utility maximizer assigns all accepted answers to the Guardian branch. Such a boundary policy can legitimately maximize empirical utility inside the certified set. Non-trivial use of both stages is an empirical outcome, not a constraint of the method.

E.12.3 Correct-answer recovery

Recovery measures what certification retains: accepted correct answers divided by test examples on which at least one cached model is correct, computed within each split. Higher recovery alone does not imply a safer router; safety concerns the separate population certification event.

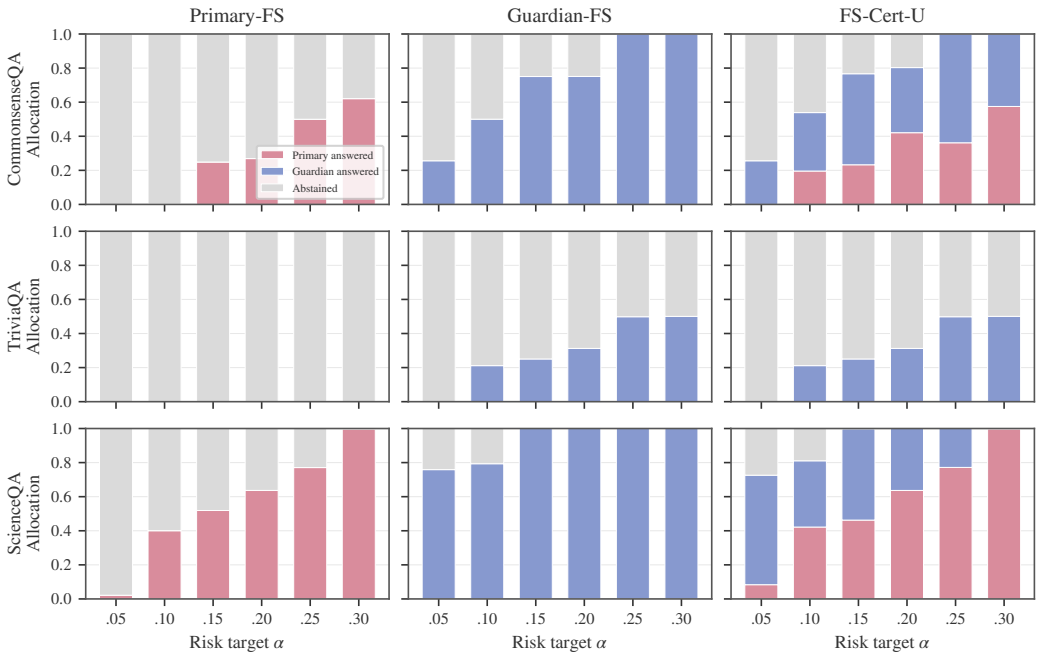


Figure 28: Routing allocation for Qwen3.5-2B→Qwen3.5-9B. Each bar partitions all test questions into Primary answered, Guardian answered, and abstained fractions, averaged over 500 matched splits. Columns compare separately certified single-model controls and FS-Cert-U; Guardian invocation may exceed the Guardian answered fraction.

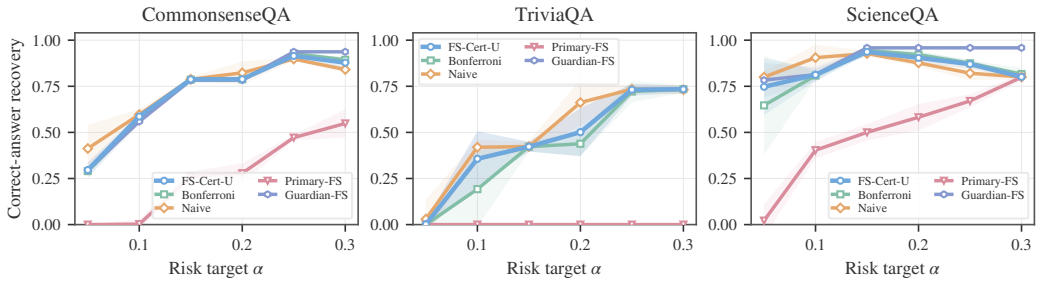


Figure 29: Correct-answer recovery for the same pair: accepted correct answers divided by questions for which either cached model is correct, computed within each split. Lines and bands are the mean and standard deviation over 500 splits, including zero-output runs. Naive lacks certification; higher recovery alone does not establish a safer router.

E.12.4 Cost-weight sensitivity

We rerun screening tie-breaks and utility selection at each prespecified $\lambda \in \{0, 0.10, 0.25, 0.50\}$, keeping $\alpha = 0.10$, candidate families, and all 500 splits fixed. The $\lambda = 0.25$ replay exactly reproduces the original answered, error, call, and utility records. Main-text Figure 7 reports utility averaged over all 11 pairs per dataset. The FS–Bonferroni utility advantage persists at every weight; it ranges from 0.0207–0.0398 on CommonsenseQA, 0.0616–0.0755 on TriviaQA, and 0.1048–0.1136 on ScienceQA. Naive has higher utility, but no set-level risk certificate. At $\lambda = 0.50$, even FS utility is negative on TriviaQA (-0.1404): invocation cost can outweigh retained coverage. Certification cannot manufacture a favorable operating point when Guardian invocation is sufficiently expensive relative to attainable coverage. Appendix Figure 30 reports the corresponding discrete call-rate–coverage operating points; these are not a continuous frontier or equal-budget comparisons. The positive FS–Bonferroni differences support robustness of the fixed-sequence certification-power advantage to the cost weight. We preserve the original routed all-abstain fallback, which still charges a Guardian call, rather than silently replacing empty runs by zero-cost outcomes.

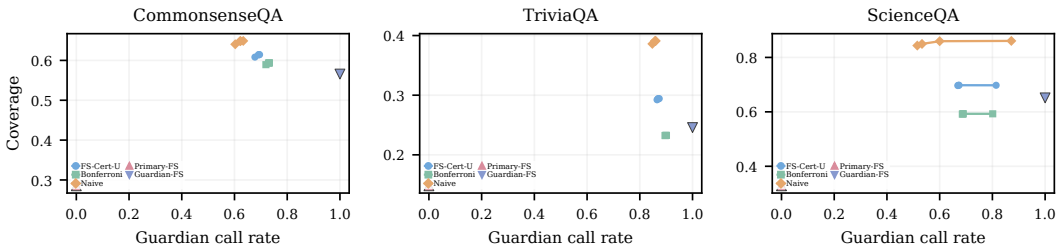


Figure 30: Guardian-call rate versus coverage operating points at fixed $\alpha = 0.10$, connected in increasing $\lambda \in \{0, 0.10, 0.25, 0.50\}$. Primary-FS and Guardian-FS references are shown. These are discrete operating points, not a continuous frontier or a comparison at identical call budgets.

E.12.5 Post-certification utility selection

To complement, rather than replace, the FWER replay, we freeze a separate two-policy, three-type population with exact SCER 0.02 for both policies and utilities 0.875 and 0.8625. The policies trade Primary acceptance against Guardian answering on different input types, so neither dominates pointwise. At $(\alpha, \delta, \lambda) = (0.10, 0.05, 0.25)$, 5,000 independent repetitions per certification size yield mean population regret 0.003465, 0.002635, 0.001535, 0.000678, and 0.0001525 for $m = 280, 560, 1120, 2240, 4480$. The probability of positive regret falls from 27.72% to 1.22%. Every run returns a nonempty certified set and an actually safe selected policy in this stress population. The theoretical $2\varepsilon_U$ decreases from 0.2211 to 0.0553 and remains conservative; no observed population regret exceeds it. This supports the qualitative finite-sample behavior predicted by Theorem 2: near-tied certified utilities permit finite-sample utility mis-selection, and this effect decreases as certification size increases. The experiment supports the theorem's predicted sample-size trend, rather than its numerical sharpness; it does not prove the theorem experimentally. Full population specification and seeds are given in Appendix E.9.

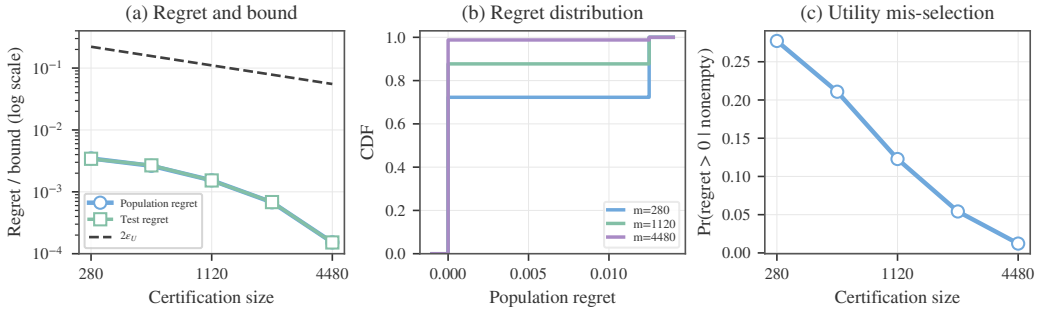


Figure 31: Near-tie utility stress test, independent of the original FWER experiment. (a) Mean certified-set population and independent-test regret, with $2\varepsilon_U$ for $K = 2$, $\eta = 0.05$, on a common logarithmic scale. Independent test size is $20m$. (b) Regret CDFs conditional on nonempty certification. (c) Probability of selecting a suboptimal policy within the certified set. Exact population utilities, not a finite oracle pool, define population regret.

F Extended Related Work

F.1 Selective prediction and abstention

Selective prediction augments a predictor with a reject option and evaluates the resulting coverage–risk trade-off. Classical selective classification formalizes the risk of accepted examples and studies how a confidence rule determines which examples are retained (Geifman & El-Yaniv, 2017). SelectiveNet incorporates the coverage constraint into representation learning, making abstention part of the predictive objective rather than a post-hoc threshold (Geifman & El-Yaniv, 2019). These formulations provide the conceptual antecedent for our answered indicator S_θ: abstention changes the population on which error is measured, so reliability cannot be summarized by unconditional model accuracy alone. In our setting the reject option is embedded in a two-stage policy. A query can be answered by the Primary, escalated to the Guardian, or abstained, and the accepted population depends jointly on both thresholds. FS-Cert-U uses selective-prediction language for this system-level event, while keeping the underlying model outputs and uncertainty scores fixed during certification.

F.2 Uncertainty quantification and conformal reliability for LLMs

Semantic uncertainty methods cluster generations or measure invariance across equivalent answers (Kuhn et al., 2023; Farquhar et al., 2024); sequence-level entropy and related scores provide another way to rank uncertainty in free-form question answering (Wang et al., 2025b). These estimators are useful routing signals, but ranking confidence does not itself specify a finite-sample error guarantee. Conformal prediction instead calibrates scores or sets to obtain distribution-free coverage statements (Angelopoulos & Bates, 2021; Quach et al., 2024), and ConU transfers this idea to correctness coverage for LLM outputs (Wang et al., 2024). SConU adds selective conformal uncertainty for retaining a subset of outputs (Wang et al., 2025c). Other recent methods target related but distinct events: COIN studies uncertainty-guarded selective question answering with provable risk control (Wang et al., 2026b); SAFER constrains risk in a sample-then-filter generation pipeline (Wang et al., 2025a); and MiRD decomposes miscoverage for reliable set-valued open-ended QA (Hu et al., 2026). LEC develops selection-conditioned risk control for selective prediction and two-model routing (Wang et al., 2026a). Thus these works span uncertainty ranking, conformal coverage, selective answering, and decomposition of open-ended QA risks; they do not all control the same event. FS-Cert-U focuses on SCER for the answered population produced by a finite two-stage policy family.

F.3 Model routing and cascades

Routing methods optimize the allocation of queries across models with different costs and capabilities. FrugalGPT combines model cascades and selection rules to reduce cost while maintaining quality (Chen et al., 2023), and RouteLLM learns a cost–quality router from preference data (Ong et al., 2025). Trust or Escalate studies escalation decisions with guarantees tied to agreement with a judge or reference decision (Jung et al., 2025). In retrieval-augmented systems, BalanceRAG calibrates risk jointly across stages of a cascade (Jia et al., 2026); LEC expresses a related selection-conditioned objective through linear expectation constraints and extends it to two-model routing (Wang et al., 2026a). These methods address complementary deployment settings and inferential targets. FS-Cert-U does not claim that their guarantees are ordered by strength. It targets a finite routing-policy family whose entire certified subset is protected before downstream utility selection, so the final optimizer can trade coverage against Guardian calls inside that protected set.

F.4 Finite-sample risk control

PAC confidence sets and distribution-free risk-controlling prediction sets provide finite-sample statements for general prediction risks (Park et al., 2020; Bates et al., 2021). Conformal Risk Control calibrates a threshold for a bounded monotone loss, while the non-monotonic extension replaces monotonicity with an algorithmic-stability condition (Angelopoulos et al., 2024; Angelopoulos, 2026). Learn-then-Test separates algorithm construction from a collection of risk tests and then uses multiple-testing logic to retain valid candidates (Angelopoulos et al., 2025). Conformal p-value selection similarly allows a data-dependent selection step with calibrated p-values (Jin & Candès, 2023). These frameworks can be applied to many policy classes, but their inferential object and calibration structure differ. FS-Cert-U certifies a finite set of routing policies directly, rather than assuming a globally monotone threshold path or certifying a continuously optimized threshold. The finite family is frozen before certification, and the SCER test conditions on the random number of answered examples through an exact binomial calculation.

F.5 Multiple testing and fixed-sequence procedures

Bonferroni and Holm procedures provide general multiplicity corrections, while fixed-sequence procedures exploit a pre-specified order to test reached hypotheses at the full level and retain strong family-wise error

control (Holm, 1979; Qiu et al., 2015; Bretz et al., 2009). Their power benefit depends on placing promising hypotheses early; the order must therefore be fixed independently of the outcomes used for certification. This logic is related to Learn-then-Test, but the routing construction has additional structure: candidate policies share the same certification sample while selecting different answered subsets, each risk denominator is random, and stopping returns a prefix that is subsequently exposed to a utility optimizer. The novelty is not fixed-sequence testing itself. It is the routing-specific construction of a frozen finite policy family, an independently learned order, an exact candidate-wise test for selection-conditioned SCER, and safe post-certification utility optimization over the admitted set. The resulting comparison is deliberately within a matched frozen family: generic corrections remain valid and may be more conservative, while FS-Cert-U seeks additional power without changing the target risk event.