TO ABSTAIN OR TO RE-RANK? A REGIME LAW FOR RISK-CONTROLLED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Risk-controlled generation returns a prediction set of candidate answers while controlling the probability that none is admissible. Existing methods calibrate sampling and filtering rules to satisfy this constraint, but the smallest prediction set permitted by the risk budget is not characterized. We study this efficiency problem directly. Classical selective prediction gives the optimal one-answer rule, and its instantiation here yields a simple regime law. Let e_0 be the Bayes top-1 error and T the target miscoverage risk. If $e_0 < T$, a valid one-answer predictor has slack and average prediction-set size is reduced by abstaining on low-confidence questions. If $e_0 > T$, no one-answer predictor is valid and efficiency depends on ranking multiple candidates. We also show that a score carrying only within-question rank cannot selectively abstain because it lacks a confidence scale across questions. These results motivate ACF-LV, a single-stage conformal filter with a calibration-gated verifier. On three QA datasets and five language models, the high-probability single-stage variant gives smaller prediction sets on 13 of 15 settings. Under a unified end-to-end protocol, ACF-LV reduces mean APSS from 4.389 to 1.889, while all eight testable predictions of the regime law agree with the observed effect of re-ranking. Code: `https://anonymous.4open.science/r/acflv-A6E8/README.md`

1 INTRODUCTION

Large language models can generate many plausible responses to the same question, but a single sampled response is not reliable enough for risk-sensitive use. Conformal methods offer a model-agnostic alternative. They sample candidate answers, construct a prediction set, and control the probability that the set contains no admissible answer (Quach et al., 2024; Mohri & Hashimoto, 2024; Abbasi Yadkori et al., 2024a;b). This changes the output of the language model from one answer to a statistically calibrated set.

Coverage is only part of the problem. A prediction set is useful when it is also small. Recent work therefore calibrates sampling budgets and filtering thresholds to control miscoverage while reducing average prediction-set size, or APSS (Wang et al., 2025a; 2026). These methods begin with a sampling or uncertainty score and then calibrate the resulting procedure. We take the opposite view. We first ask what the minimum APSS is under a fixed miscoverage budget, then ask what a practical score must preserve to approach it.

The answer is determined by a quantity that is already present before any conformal calibration. Let e_0 denote the Bayes error of returning the best candidate for every question and let T be the target miscoverage risk. If $e_0 < T$, a one-answer predictor is already valid. It can reduce APSS below one by abstaining on the least confident questions. If $e_0 > T$, no predictor that returns at most one answer can satisfy the risk constraint. Some questions must retain multiple candidates, so efficiency depends on which candidates are kept. We call these the abstention and ranking regimes. The transition occurs at $e_0 = T$.

This observation follows from a selective prediction view of answer-set filtering. Chow's reject rule gives the Bayes-optimal one-answer policy (Chow, 1970; Franc et al., 2023). Minimum expected set results characterize the corresponding coverage-size frontier (Sadinle et al., 2019). We instantiate these results for the event used in open-ended generation, where failure means that no returned

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.

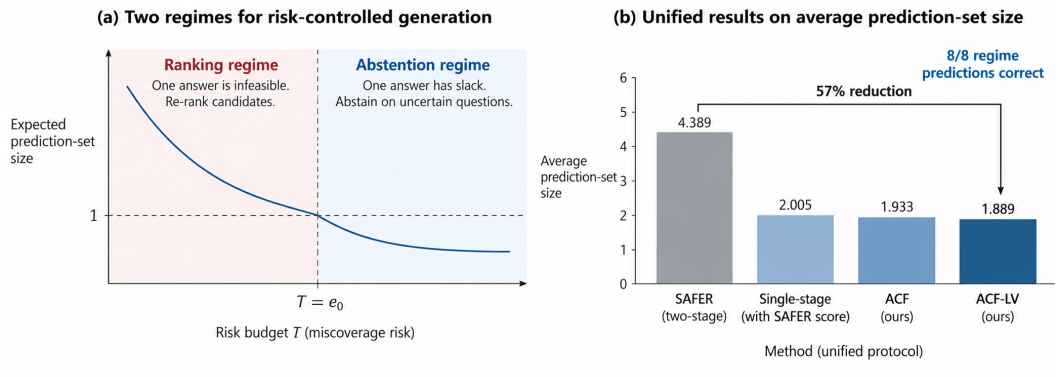

Figure 1: **Risk-controlled generation has two efficiency regimes.** The Bayes top-1 error e_0 determines whether a valid one-answer predictor has slack. When $e_0 < T$, APSS can be reduced by abstaining on uncertain questions. When $e_0 > T$, one answer is infeasible and the problem becomes candidate ranking. Under the unified evaluation in **??**, ACF-LV reduces mean APSS from 4.389 to 1.889 and all eight testable regime predictions have the expected direction.

candidate is admissible. The resulting regime law is simple, but it changes how the score should be interpreted.

A direct consequence concerns rank-only confidence scores. A score that records which candidate is first, second, and third within each question does not say whether the first candidate for one question is more reliable than the first candidate for another. That information is irrelevant when multiple answers are unavoidable, but it is exactly what selective abstention needs. We show that, under a common deterministic threshold, a within-question rank score cannot selectively return the empty set and is therefore bounded below by one answer whenever its one-answer rule is valid. Better ranking can help in the ranking regime without solving the abstention problem.

These results motivate ACF-LV. The first change is to calibrate the final miscoverage event directly rather than allocate separate risk budgets to sampling and filtering. The second is a nonconformity score that is rank-first within a question but retains a small cross-question entropy term. The third is a learned verifier that is used only when a calibration diagnostic indicates that the entropy ranking is poor. ACF-LTT adds Learn-then-Test certification over the calibration choices (Bates et al., 2021; Angelopoulos et al., 2025; 2026). The method does not require fine-tuning the language model.

The experiments use the three QA datasets and five language models in the SAFER benchmark. The main results are reported in the paper rather than deferred to the appendix. ACF-LTT reduces APSS in 13 of 15 settings. Under a unified protocol with identical train, calibration, and test splits, ACF-LV reduces mean APSS from 4.389 to 1.889. A controlled ablation shows that single-stage calibration accounts for 95.4% of this reduction. All eight testable regime predictions have the expected direction.

Our contributions are threefold. We characterize APSS under a fixed miscoverage budget for the event that at least one returned answer is admissible. The resulting regime law separates abstention from re-ranking through the comparison between e_0 and T. We show that within-question rank alone cannot support selective abstention under a common deterministic threshold. This identifies cross-question confidence as the information needed to obtain APSS below one. We introduce ACF-LV and ACF-LTT and evaluate them on three datasets and five language models. The main paper reports the full 15-setting benchmark, the end-to-end ablation, the regime test, and the principal robustness results.

2 Related Work

Risk-controlled generation. A growing line of work uses conformal methods to control errors in language generation. Some methods calibrate how many candidates to sample (Quach et al., 2024; Wang et al., 2025a), while others use uncertainty to filter or abstain (Mohri & Hashimoto, 2024; Abbasi Yadkori et al., 2024a;b). SAFER combines an abstention-aware sampling stage

with conformal filtering and provides the benchmark suite used in our experiments (Wang et al., 2026). These approaches build on split conformal prediction and risk-controlling prediction sets (Papadopoulos et al., 2002; Lei et al., 2018; Vovk et al., 2022; Angelopoulos et al., 2024). Learn-then-Test provides the high-probability certification used in our stronger finite-sample variant (Bates et al., 2021; Angelopoulos et al., 2025; 2026).

Selective prediction. Selective prediction studies when a model should answer and when it should abstain. The Bayes-optimal reject rule is due to Chow (1970), with modern characterizations of bounded abstention and reject-option learning in Franc et al. (2023) and distribution-free variants in Szabadváry et al. (2025). Risk-coverage curves summarize the resulting trade-off (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Selective question answering applies the same principle to language tasks (Kamath et al., 2020). Our theory instantiates this view for multi-candidate answer sets under the event that at least one returned answer is admissible.

Efficiency of conformal sets. The closest theoretical foundation is the minimum expected set-size result of Sadinle et al. (2019). We use it to express the one-answer optimum as a functional of the confidence distribution. Other work studies tighter conformal sets through adaptive scores and efficiency criteria (Romano et al., 2020; Angelopoulos et al., 2021; Vovk et al., 2016; Dhillon et al., 2024). Xi et al. (2025) show that ranking rather than score calibration drives set-size efficiency when prediction sets are nonempty. Our setting permits abstention. This changes the role of ranking because a score that preserves only within-question order loses the cross-question scale needed to abstain selectively. Concurrent work studies singleton conformal sets and combinations of risk control with selective prediction (Wang et al., 2025b; Xu et al., 2025; Bai & Jin, 2026). Additional discussion appears in Section J.

3 THEORY

We now solve the size minimization problem behind risk-controlled answer sets. The key step is to view filtering as selective prediction. This gives an optimal score, an efficiency limit, and the boundary that separates abstention from re-ranking. Full proofs are given in Section A.

3.1 SETUP

A question $X \sim P_X$ produces m candidates $(Y_j)_{j=1}^m$ from a frozen language model. Candidate j has an admissibility label $A_j \in \{0, 1\}$ and reference-free features $\Phi_j = \phi(X, Y_j)$. The labels are observed during calibration and hidden at test time. A score $s_j = s(\Phi_j)$ is smaller for candidates we prefer to keep. For threshold t we return

$$C_t(X) = \{j \mid s_j \le t\}, \qquad c(X) = \min\{s_j \mid A_j = 1\}, \tag{1}$$

where $c(X) = +\infty$ if no candidate is admissible. Failure occurs when $c(X) > t$. We therefore control

$$R(t) = \mathbb{P}(c(X) > t) \le T \tag{2}$$

and minimize the expected set size $N(s) = \mathbb{E}|C_t(X)|$.

Write $q_j(X) = \mathbb{P}(A_j = 1 \mid \Phi_j)$ for the correctness posterior and $p(X) = \max_j q_j(X)$ for the confidence of the best candidate. Let F_p be the distribution of p and let $e_0 = \mathbb{E}[1 - p(X)]$ be the Bayes top-1 error.

Assumptions. We use four assumptions locally rather than globally. A0 requires coverage feasibility, so $\mathbb{P}(\exists j\ A_j = 1) \ge 1 - T$. A1 is exchangeability of calibration and test examples. A2 is conditional independence of $(A_j)_j$ given $(\Phi_j)_j$ and is needed only for the exact multi-candidate optimum. It is not needed for filters that return at most one answer. A3 assumes an atomless coverage score when a continuous population threshold is used. The appendix states each assumption again where it enters a proof.

3.2 VALIDITY AND EFFICIENCY

Lemma 1 (Split-conformal validity)**.** *Under A1, for any score s that does not depend on calibration or test labels, let $c_{(1)} \le \cdots \le c_{(n)}$ be the ordered calibration coverage scores and set $\hat{t} = c_{(k)}$ with $k = \lceil (n+1)(1-T) \rceil$. Then*

$$T - \frac{1}{n+1} \le \mathbb{P}(c(X_{\text{test}}) > \hat{t}) \le T,$$

with $\hat{t} = +\infty$ when $k > n$.

The validity statement does not depend on which fixed score is used. The score changes efficiency, not the conformal logic that controls risk.

Lemma 2 (Monotone invariance)**.** *For every strictly increasing function ψ, the score $\psi \circ s$ gives the same conformal set and the same expected size N. Thus $N(s)$ depends on the joint order induced by the score across candidates and questions. It is not determined by AUC alone. Two equal-AUC scores with different values of N are constructed in Section A.*

This distinction matters because a good within-question ranking need not preserve the confidence scale across questions. The latter becomes essential once abstention is allowed.

3.3 THE OPTIMAL FILTER AND ITS LIMIT

Theorem 1 (Chow optimality, borrowed and instantiated here)**.** *Assume A0 and minimize $\mathbb{E}|S(X)|$ over all measurable, possibly randomized filters subject to $\mathbb{P}(S$ covers no admissible answer$) \le T$.*

One-answer filters. *On $\mathcal{S}_1$, the class of filters that return at most one answer per question, the size-optimal rule commits to $\arg\max_j q_j$ when $\max_j q_j \ge \theta_T$ and abstains otherwise, with randomization at the budget boundary. The optimal score is $s_j = -q_j$. Conformal calibration recovers the corresponding threshold and the achievable size-risk frontier is the risk-coverage curve of $p = \max_j q_j$.*

Multi-answer filters. *Under A2, the optimum adds candidates in decreasing q_j while $q_j \prod_{k \in S}(1 - q_k) \ge 1/\lambda_T$. Without A2 we claim only the $(1 - 1/e)$ approximation stated and proved in Section A.*

The first part is classical reject-option optimality (Chow, 1970; Franc et al., 2023; Szabadváry et al., 2025). Our use of it is the instantiation under the event that a returned set contains at least one admissible answer. That step identifies the relevant posterior and makes the efficiency limit explicit.

Corollary 1 (Efficiency limit, from Sadinle et al. (2019))**.** *Assume A0 and $e_0 \le T$. Let p_τ satisfy*

$$\int_{p \le p_\tau} p \, \mathrm{d}F_p(p) = (T - e_0)_+$$

and let $G(T) = F_p(p_\tau)$ be the mass of questions on which the optimal one-answer filter abstains. The minimum expected set size over $\mathcal{S}_1$ is

$$N_{\text{Bayes}}(T) = 1 - G(T) \in [1 - T, 1]. \tag{3}$$

Any valid threshold filter has $N \ge 1 - T$. Equality requires the budget to be spent entirely on zero-confidence mass and may require randomization at an atom.

The interpretation is simple. Abstaining on a question with confidence p saves one unit of set size at a risk cost of p. The optimum spends the available slack $T - e_0$ on the least confident questions first. A second correct paraphrase has no value for this objective once one admissible answer is already present.

3.4 THE REGIME LAW

Theorem 2 (Regime law, new)**.** *Assume A0 and $\mathbb{P}(p = 0) = 0$. Then*

$$\begin{aligned} &N_{\text{Bayes}}(T) < 1 \iff e_0 < T, \qquad N_{\text{Bayes}}(T) = 1 \iff e_0 = T, \\ &\mathcal{S}_1 \text{ is infeasible} \iff e_0 > T. \end{aligned} \tag{4}$$

When $e_0 < T$, efficiency is obtained by abstaining on the least confident questions. When $e_0 > T$, no one-answer filter satisfies the risk budget and every valid solution must keep multiple candidates on a set of positive probability. The proof follows from one reference policy. Returning the best candidate everywhere has size exactly one and risk exactly e_0. Whether that policy has slack determines which side of the boundary the problem occupies.

The regime law also clarifies what information a useful score must retain.

Corollary 2 (Scale-free scores cannot abstain, new)**.** *Call a score scale-free when it is a fixed increasing function of within-question rank alone and therefore carries no cross-question confidence scale. Fix $T < 1$, a deterministic threshold, and a common candidate count m. Then the coverage score lies on the same m-point rank grid for every question. The returned set size is the same integer for every question and selective abstention is impossible. Whenever the one-answer rule is valid, the smallest valid deterministic choice has $N = 1$ at realized risk $1 - \pi_0$, where π_0 is the probability that the top-ranked candidate is admissible. Hence a scale-free score is strictly above $N_{\mathrm{Bayes}}(T)$ in the abstention regime.*

A rank-only score can still be an excellent re-ranker. The limitation is narrower. Once the score discards cross-question scale, a common deterministic threshold cannot identify which questions should receive no answer. This reverses the usual emphasis on ranking for conformal efficiency (Xi et al., 2025) because empty sets are part of the decision space here.

Finite-sample guarantees. Lemma 1 covers a fixed score. A learned score can be fitted on a training fold disjoint from calibration and test. Selecting an operating point from a finite grid requires a separate high-probability argument. We use Learn-then-Test for that purpose (Bates et al., 2021; Angelopoulos et al., 2025). Details are given in Section A.7.

Scope of novelty. The selective prediction equivalence and the minimum expected set result are prior work. The new theoretical claims are Theorem 2 and Corollary 2. Two related questions remain open. The first is a scalar closed form for the ranking-regime optimum without the conditional-independence simplification. The second is a strict minimax lower bound over a restricted feature class.

4 METHOD

The theory suggests two design principles. The final answer set should be calibrated directly against the user risk budget, and the score should preserve cross-question confidence whenever abstention can help. ACF-LV follows both principles.

Single-stage calibration. For each calibration question we compute the coverage score $c(X) = \min\{s_j \mid A_j = 1\}$ from Equation (1). A single threshold is then calibrated at the final risk level T. This differs from a two-stage construction that first controls a sampling event at level α and then controls filtering at level β, with the two levels composed as $T = \alpha + \beta - \alpha\beta$. The second construction pays for two calibrated events even though the user ultimately cares about one event, whether the returned set contains an admissible answer.

For a fixed score, ordinary split conformal calibration gives the marginal variant ACF. We also use a finite grid of conservative margins κ and certify the selected operating point with Learn-then-Test. This gives ACF-LTT, our high-probability variant. With ordered coverage scores $c_{(1)} \le \cdots \le c_{(n)}$, the candidate threshold is

$$\hat{t} = c_{(k)}, \qquad k = \lceil (n+1)(1 - T_\kappa) \rceil - 1, \tag{5}$$

where

$$T_\kappa = \max\left(T - \frac{1-T}{n+1} - \kappa \sqrt{\frac{T(1-T)}{n}}, 0 \right).$$

The margin grid is selected without using test labels. Learn-then-Test certifies the selected configuration at a stated confidence level (Bates et al., 2021; Angelopoulos et al., 2025).

Table 1: **Unified end-to-end comparison.** All arms use the same three-way splits and the same 100 reshuffles. The share of reduction is measured relative to the total decrease from SAFER to the complete ACF-LV method. PAC-SAFER replaces the second calibration stage with Learn-then-Test and is reported separately because it strengthens the guarantee rather than forming part of the ablation ladder.

Method	Mean set size	Per-split validity	Share of reduction
SAFER	4.389	79.4%	reference
Single-stage, SAFER score	2.005	100.0%	95.4%
+ rank score (ACF)	1.933	100.0%	2.9%
+ gated verifier (ACF-LV)	1.889	100.0%	1.8%
PAC-SAFER	5.257	99.9%	matched HP baseline

A score that can abstain. The Bayes score $-q$ is unavailable at test time, so we use length-normalized predictive token entropy u_j as a reference-free proxy. Pure within-question rank is attractive because it is robust to question-specific entropy scale, but Corollary 2 shows that it removes the cross-question information needed for selective abstention. We therefore use

$$s_j = \mathrm{rank}_j(u) + \epsilon z_j, \tag{6}$$

where the rank is normalized within each question and $z_j \in [0, 1]$ is a min-max normalization of entropy across the evaluation cell. We set $\epsilon = 0.05$. With $m = 20$, the rank gap is $1/(m-1) = 0.0526$. The chosen value is below this gap, so z_j restores a cross-question scale without changing the within-question rank order. The sensitivity experiment in Section D confirms that every tested value below the gap gives effectively identical results. The constant is therefore structural rather than tuned for performance.

Verifier fallback. In the ranking regime, one answer is not enough and better candidate ordering can reduce the number of answers that must be retained. We fit a gradient-boosted LambdaRank verifier (Ke et al., 2017) using reference-free features listed in Section B. The verifier is trained on a fold that is disjoint from calibration and test, which preserves the validity argument conditional on the training fold. Its output is converted to within-question rank. This makes the verifier useful for ordering candidates but unable to exploit selective abstention on its own.

The verifier is used only when a calibration diagnostic suggests that entropy is a poor ranking signal. We switch it on when

$$\mathrm{AUC}(-u, \mathbf{1}[A = 1]) < \frac{1}{2}. \tag{7}$$

This gate is a heuristic proxy for the ranking regime rather than a consequence of Theorem 2. It uses calibration labels and never test set size. The unified experiments in Section 5.4 evaluate the complete gated procedure on the same splits as every comparison arm. Algorithm 1 gives the full algorithm.

5 Experiments

The experiments ask two questions. First, does calibrating the final answer set directly reduce set size at the same or stronger level of reliability? Second, does the role of re-ranking change at the one-answer boundary predicted by Theorem 2?

5.1 Setup

We use ScienceQA (Lu et al., 2022), TriviaQA (Joshi et al., 2017), and CoQA (Reddy et al., 2019). Each dataset is crossed with Qwen2.5-7B, Llama-3.1-8B, Qwen2.5-3B, Qwen2.5-14B, and OpenChat-3.5 (Qwen et al., 2024; Grattafiori et al., 2024; Wang et al., 2024). This gives the same 15-cell text benchmark grid used by Wang et al. (2026). We regenerate candidates with the released pipeline at its published defaults and evaluate all methods on identical candidate pools. Each question has at most $m = 20$ sampled candidates. The default admissibility rule is reference similarity of at least 0.6. The sampling level is $\alpha = 0.05$ on ScienceQA and TriviaQA and $\alpha = 0.25$ on CoQA. We sweep $\beta \in \{0.05, 0.10, \ldots, 0.50\}$ and report the corresponding combined risk $T = \alpha + \beta - \alpha\beta$.

The unified evaluation uses one estimator implementation for every arm. Each reshuffle assigns 0.3 of the data to verifier training, 0.2 to calibration, and 0.5 to test. We use 100 reshuffles and identical

Table 2: Prediction-set efficiency on the 15-setting benchmark. APSS is averaged over 100 calibration-test reshuffles and ten risk levels. Valid is the percentage of reshuffles whose test EER is at most the target risk. ACF-LTT uses Learn-then-Test with $\delta = 0.1$.

Dataset	Backbone	SAFER		ACF-LTT		Δ APSS
		APSS	Valid	APSS	Valid	
ScienceQA	Qwen2.5-7B	0.739	82%	0.794	100%	+7%
	Llama-3.1-8B	1.440	76%	0.876	100%	-39%
	Qwen2.5-3B	1.386	77%	0.815	99%	-41%
	Qwen2.5-14B	0.729	82%	0.793	100%	+9%
	OpenChat-3.5	1.296	79%	0.808	100%	-38%
TriviaQA	Qwen2.5-7B	3.291	77%	0.950	100%	-71%
	Llama-3.1-8B	3.180	81%	1.039	100%	-67%
	Qwen2.5-3B	5.003	73%	1.426	100%	-71%
	Qwen2.5-14B	1.983	81%	0.892	100%	-55%
	OpenChat-3.5	7.296	68%	1.302	100%	-82%
CoQA	Qwen2.5-7B	9.985	80%	3.179	100%	-68%
	Llama-3.1-8B	1.186	99%	0.854	100%	-28%
	Qwen2.5-3B	7.392	84%	2.889	100%	-61%
	Qwen2.5-14B	10.292	71%	3.249	100%	-68%
	OpenChat-3.5	1.385	99%	0.880	100%	-36%

folds across all arms. This makes the entropy and verifier comparisons directly comparable. The implementation also reproduces the original comparison values exactly on representative cells, as shown in Table 5.

5.2 Benchmark results on all 15 settings

Table 2 reports the complete benchmark. ACF-LTT returns a smaller prediction set on 13 of 15 dataset-model settings. The median APSS reduction over those 13 settings is 60.9%, or 55.0% over all 15 settings. The two exceptions are ScienceQA with Qwen2.5-7B and Qwen2.5-14B, where both methods already operate below one answer on average. Across the 15 settings, the fraction of reshuffles with test EER at or below the target is 80.5% for SAFER and 99.8% for ACF-LTT.

5.3 End-to-end efficiency

Table 1 gives the main result. The complete ACF-LV method returns a mean set size of 1.889 compared with 4.389 for the two-stage reference procedure. This is a 57% reduction. Observed per-split validity is 100.0% for ACF-LV and 79.4% for the reference procedure. The absolute set sizes are larger than in the 0.5 calibration split because the unified protocol reserves data for verifier training. The ordering is unchanged.

A stronger comparison replaces the second stage of the two-stage procedure with Learn-then-Test. This PAC-SAFER variant reaches 99.9% observed per-split validity and has mean set size 5.257. ACF-LV retains a 64% reduction against this stronger baseline while reaching 100.0% observed per-split validity. We also repeat our high-probability calibration at delta 0.05. Mean set size increases by only 2.7% relative to delta 0.1, with 99.9% observed per-split validity and 99.9% certification rate. Full values appear in Section F.

Where the gain comes from. The ablation in Table 1 gives a clear answer. Applying single-stage calibration to the baseline score reduces mean set size from 4.389 to 2.005. Relative to the total reduction from 4.389 to 1.889, this step accounts for 95.4% of the gain. Replacing the raw score with the rank-based score contributes another 2.9%, and the gated verifier contributes the remaining 1.8%. The large headline gain therefore comes from calibrating the final objective directly, not from a stronger learned score. The rank transform and verifier matter mainly because they let us test and exploit the two regimes predicted by the theory.

Table 3: **The regime prediction under the unified protocol.** Entropy set size is measured before fitting the verifier. Below one answer the rank-only verifier is predicted to hit its one-answer floor. Above one answer re-ranking is the available source of improvement. All eight testable directions agree with the observation.

Dataset	Backbone	Entropy set size	Prediction	Observation
SciQA	Qwen2.5-7B	0.938	floor	floor
SciQA	Qwen2.5-3B	1.347	reduce	reduce
SciQA	Qwen2.5-14B	0.857	floor	floor
TriviaQA	Qwen2.5-14B	0.990	floor	floor
CoQA	Qwen2.5-7B	4.758	reduce	reduce
CoQA	Qwen2.5-3B	4.598	reduce	reduce
CoQA	Qwen2.5-14B	4.941	reduce	reduce
CoQA	OpenChat-3.5	1.019	reduce	reduce

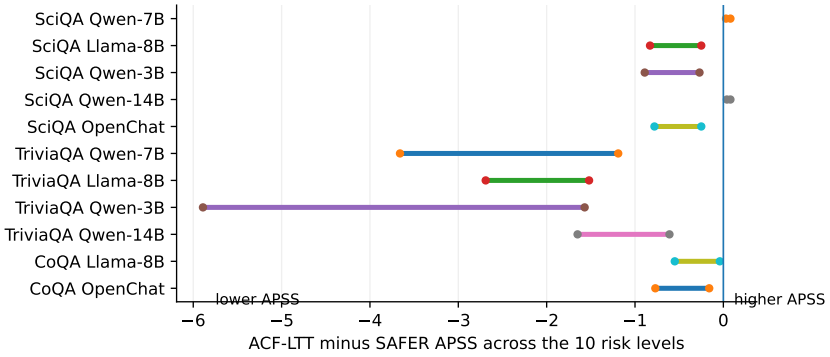

Figure 2: **Difference in APSS across the ten risk levels.** Each segment gives the smallest and largest value of ACF-LTT APSS minus SAFER APSS for one pairable dataset-model setting. Negative values favour ACF-LTT. Nine of the eleven intervals lie entirely below zero. Four additional settings are omitted because SAFER is infeasible at one or more risk levels.

5.4 The regime law under one protocol

Table 3 removes the earlier protocol confound by evaluating the entropy score and verifier on the same folds. The result is exact at the level of direction. All eight testable cells behave as predicted. The three cells below one answer hit the rank-only floor. The five cells above one answer are reduced by re-ranking. This includes ScienceQA with Qwen2.5-3B as well as the CoQA cases, so the effect is not confined to a single dataset-model combination. The calibration gate is intentionally conservative. Three TriviaQA cells have entropy AUC above 0.5, so the gate declines to fit a verifier. We force a verifier on these cells as a specificity check. Set size increases on all three by 0.4% to 5.9%. On this grid the gate therefore has no observed false negative in the sense of missing a helpful verifier. The full table is given in Section G.

5.5 Robustness

The reduction is not an artifact of averaging over risk levels. On 9 of the 11 cells for which every risk level is pairable, the method returns a smaller set at all ten risk levels. Across the 110 paired cell-risk points, 109 differences are significant at the 99% level over the 100 matched reshuffles. The two cells without a gain are ScienceQA settings where both procedures already return fewer than one answer. Four additional cells are excluded from the paired count because the baseline is infeasible at some risk levels. They include several of the largest reductions. The per-cell summary appears in Section E. The tie-break in Equation (6) is also stable. At $m = 20$ the structural rank gap is $1/(m-1) = 0.0526$. Sweeping ϵ from 0.01 through the boundary changes mean set size by less than 10^{-4} at the cell level. Once ϵ moves above the gap, all cells can change because the score is no longer rank-first. The sweep in Section D supports the intended interpretation of $\epsilon = 0.05$ as a structural choice rather than a tuned value.

Table 4: Robustness to the admissibility criterion. Strict wins require smaller APSS and no lower observed per-split validity than SAFER. Infeasible cells are excluded from the win count because SAFER does not control risk there.

Criterion	Feasible	Strict wins	Median APSS reduction	Infeasible
Similarity	15	13	55%	0
ROUGE-L	11	7	25%	4
Entailment	12	7	21%	3
LLM judge	11	8	31%	4
Total	49	35	–	11

5.6 Other correctness criteria

We repeat the original 0.5 calibration analysis under all four correctness criteria used by Wang et al. (2026). Across the 49 cells where the two-stage procedure is feasible, ACF-LTT has both smaller sets and at least as high observed per-split validity in 35 cases. The count rises to 45 when a two percentage point validity tolerance is allowed. The median set-size reduction ranges from 21% to 55% across criteria. We also vary the calibration fraction from 0.5 to 0.1. The advantage narrows as calibration data become scarce but does not reverse, with smaller sets on 14 of 15 cells at the smallest split. These results are reported in Section I.

6 Discussion and limitations

The verifier should not be read as a universal performance module. Under marginal calibration its effect can be much larger, but under the high-probability protocol used its incremental gain is about 5% on the matched verifier comparison. We use the unified experiment as the primary comparison and retain the marginal result only as a diagnostic in the appendix. There is also a generation-cost trade-off. Both procedures are evaluated on the same cached pool of 20 candidates, but their calibrated deployment budgets differ. Mean sampling budgets are similar overall, at 6.26 for the reference procedure and 6.38 for our method. The pattern varies substantially by cell. ACF-LV is much cheaper on the three CoQA Qwen cells, where the reference procedure requires roughly 13 to 16 candidates and our method requires about 3. It is more expensive on several ScienceQA cells. Smaller returned sets are therefore not obtained at uniformly lower generation cost. We report the full comparison in Section H.

The regime variable e_0 is not observed directly. Any empirical top-1 error is an upper bound on the Bayes error, so the experiments are strongest as tests of qualitative direction rather than estimates of distance from the boundary. Near $e_0 = T$, finite-sample effects can also make the regime difficult to identify. Finally, the exact multi-candidate optimum in Theorem 1 uses conditional independence of candidate admissibility. Correlated paraphrases need not satisfy that assumption. Our main one-answer regime law does not depend on it. Two theoretical questions remain open. A closed scalar expression for the ranking-regime optimum would make the boundary easier to estimate from data, and a minimax lower bound over restricted feature classes would clarify the price of using imperfect reference-free scores. We leave both questions for future work.

7 Conclusion

Risk-controlled generation is not one efficiency problem. It has an abstention regime and a ranking regime, separated by whether the Bayes top-1 error lies below or above the allowed risk. This view explains why within-question ranking alone can be insufficient and why a cross-question confidence scale becomes essential when the optimal set size falls below one. ACF-LV turns this structure into a single-stage conformal procedure with a gated verifier. The experiments show that direct calibration provides most of the efficiency gain, while the verifier behaves according to the regime predicted by the theory. The resulting picture is simple. When the risk budget leaves slack, spend it on abstention. When one answer is no longer valid, spend effort on re-ranking.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and the planning of additional ablation and robustness experiments. They were not used to generate experimental measurements. All AI-assisted text and experimental suggestions were reviewed by the authors. Numerical results were produced by the authors' analysis code. The authors checked the mathematical statements, citations, and reported values and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The submission includes the full mathematical proofs in Section A, the filtering and calibration details in Section B, and the expanded experimental tables in the remaining appendix. The experiments use fixed candidate generations for the same 15 dataset-model cells across comparison arms. The unified evaluation uses identical train, calibration, and test folds for every method and 100 reshuffles. The appendix also reports estimator validation, sensitivity to epsilon and delta, paired risk-level results, correctness-criterion robustness, calibration-size robustness, gate specificity, and sampling budgets. Code and result files needed to reproduce the reported tables and figures are released in the anonymous code link shared in the abstract.

ETHICS STATEMENT

Risk control in this work concerns an operational admissibility criterion rather than an absolute statement of truth. A guarantee at level T is an average statement under exchangeability and does not ensure correctness on every question or subgroup. The marginal and high-probability forms of the guarantee are therefore reported separately. Abstention can also concentrate on difficult or underrepresented inputs, so subgroup risk and abstention rates should be audited before deployment. Distribution shift, changes to prompting or retrieval, and updates to the underlying model can invalidate calibration and require recalibration. The experiments use public question-answering benchmarks and frozen language models. No language model is trained or released as part of this work.

REFERENCES

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024a. arXiv:2406.02543.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating LLM hallucinations via conformal abstention. arXiv preprint arXiv:2405.01563, 2024b.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2208.02814.

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; arXiv:2110.01052.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2026. To appear; prepublication version arXiv:2411.11824.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021. Spotlight; arXiv:2009.14193.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. arXiv preprint arXiv:2603.24704, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. Article 43; DOI 10.1145/3478535; arXiv:2101.02703.

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

Guneet S. Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1549–1557, 2024. arXiv:2306.07254.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Vojtěch Franc, Daniel Průša, and Václav Voráček. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. arXiv:2101.12523.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4878–4887, 2017. arXiv:1705.08500.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, et al. The Llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Ying Jin and Emmanuel J. Candes. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

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

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Pan Lu, Swaroop Mishra, Tony 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 (NeurIPS)*, volume 35, 2022. ScienceQA; arXiv:2209.09513.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047, 2024. arXiv:2402.10978.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002.

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 (ICLR)*, 2024. arXiv:2306.10193.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL `https://arxiv.org/abs/2412.15115`.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. arXiv:1808.07042.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. DOI 10.1080/01621459.2017.1395341; arXiv:1609.00451.

Johan Hallberg Szabadváry, Tuwe Löfström, Ulf Johansson, Cecilia Sönströd, Ernst Ahlberg, and Lars Carlsson. Classification with reject option: Distribution-free error guarantees via conformal prediction. *Machine Learning with Applications*, 20:100664, 2025. DOI 10.1016/j.mlwa.2025.100664; arXiv:2506.21802.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, 2012.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications (COPA)*, volume 9653 of *Lecture Notes in Computer Science*, pp. 23–39. Springer, 2016. DOI 10.1007/978-3-319-33395-3_2; arXiv:1603.04416.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2nd edition, 2022.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. OpenChat: Advancing open-source language models with mixed-quality data. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.11235.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2025a. TRON; arXiv:2410.08174.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.10193; OpenReview kJmLmOvwLC.

Tao Wang, Yan Sun, and Edgar Dobriban. Singleton-optimized conformal prediction. arXiv preprint arXiv:2509.24095, 2025b.

Huajun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. Does confidence calibration improve conformal prediction? *Transactions on Machine Learning Research*, 2025. OpenReview 6DDaTwTvdE; arXiv:2402.04344 (2024).

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. arXiv preprint arXiv:2512.12844, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

A PROOFS

What the two-sided result assumes. Corollary 2 assumes a deterministic threshold, and that is load-bearing: the floor $1 - T$ of Corollary 1 is reached only by a filter free to randomize inside the atom. Theorem 2 needs one hypothesis beyond (A0), no atom of zero confidence, since abstaining on never-coverable questions costs no budget. The filter's own distance from an atom is a design constant: the score adds $\varepsilon = 0.05$ times a globally normalized statistic to a per-question rank, just below the structural gap $1/(m-1) = 0.0526$ at $m = 20$; that constant buys back the cross-question comparability. The sweep in Section D confirms that values below this gap are effectively equivalent.

This appendix restates every result of Section 3 under its main-text number and proves it in full. The statements and proofs are those of the paper's own proof development, with five hypotheses *narrowed* here and marked at the point of use, none of them widened: the $1 - T$ floor of Corollary 1 is achievable only by a filter free to randomize at an atom of c; that corollary's water-filling *value* is stated over $\mathcal{S}_1$, the filters returning at most one answer per question; Theorem 2 carries the explicit hypothesis $\mathbb{P}(p = 0) = 0$ and its converses are obtained by trichotomy; and Corollary 2 is stated for scale-free scores under a deterministic threshold and a common candidate count. One further change is a *correction* rather than a narrowing, and we flag it as such: the equal-AUC witness of Section A.2 is *replaced*, the pair in the development having AUC 1 for both of its scores and the same N for both, so that it separates nothing; the two-question pair proved below is new, and its AUC, threshold, realized risk and set size are all computed there. Each proof re-invokes the numbered assumptions (A0) to (A3) of Section 3.1. Notation is that of Section 3.1 throughout, and we write $t^\star := F_c^{-1}(1 - T)$ for the population conformal threshold, well defined under (A0). The status of each result, standard, borrowed, a corollary of prior work, new, or open, is recorded in Section 3.4 and is not restated here.

A.1 VALIDITY

Restatement (Lemma 1)**.** *Fix any score s not depending on the calibration or test labels. Let $c_{(1)} \leq \cdots \leq c_{(n)}$ be the ordered calibration coverage scores and set $\hat{t} = c_{(k)}$ with $k = \lceil (n+1)(1-T) \rceil$ (and $\hat{t} = +\infty$ if $k > n$). Then under (A1),*

$$T - \tfrac{1}{n+1} \;\leq\; \mathbb{P}\big(c(X_{\text{test}}) > \hat{t}\big) \;\leq\; T.$$

Proof. By (A1) the rank of $c(X_{\text{test}})$ among the $n+1$ values $\{c(X_1), \ldots, c(X_n), c(X_{\text{test}})\}$ is uniform on $\{1, \ldots, n+1\}$. The event $\{c(X_{\text{test}}) > \hat{t} = c_{(k)}\}$ is the event that this rank exceeds k, of probability $(n+1-k)/(n+1)$. With $k = \lceil (n+1)(1-T) \rceil$ we have $(n+1)(1-T) \leq k < (n+1)(1-T)+1$, hence $n+1-k \leq (n+1)T$ and $n+1-k > (n+1)T - 1$, which is the two-sided bound. □

The bound is distribution-free and does not involve s: a score cannot break validity, it can only change efficiency.

A.2 Monotone invariance, and why AUC is the wrong functional

Restatement (Lemma 2). *For any strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$, the score $\psi \circ s$ produces the identical conformal set and the identical N. Hence $N(s)$ depends on s only through the order it induces on all candidates of all questions jointly, and not on AUC.*

Proof. Order statistics commute with strictly increasing maps, so the calibrated threshold transforms as $\hat{t} \mapsto \psi(\hat{t})$. Then $\{j : \psi(s_j) \leq \psi(\hat{t})\} = \{j : s_j \leq \hat{t}\}$ by strict monotonicity, so the kept set is unchanged pointwise and $N(\psi \circ s) = N(s)$. Consequently N can depend on s only through what monotone reparameterization preserves, which is the induced order. That order enters twice and the two roles are distinct: the law of $c(X) = \min\{s_j : A_j = 1\}$ fixes where the calibrated threshold falls, and the position of the remaining candidates relative to it fixes how many are kept. AUC, by contrast, is a global pairwise functional, the probability that an admissible candidate outscores an inadmissible one across all pairs, and it constrains neither role: it counts which within-question comparisons a score gets right and is blind to where in the global range they fall. So equal AUC does not determine N. □

A two-question witness (equal AUC, different N). Take two questions of equal mass, each with one admissible and one inadmissible candidate, and write each question as (admissible, inadmissible). Score s_A assigns $(0.1, 0.2)$ and $(0.8, 0.6)$; score s_B assigns $(0.2, 0.1)$ and $(0.6, 0.8)$, that is s_A with the two candidates of each question exchanged. Both score one question correctly and one incorrectly, so both have AUC $= 1/2$: the admissible values $\{0.1, 0.8\}$ of s_A outscore the inadmissible $\{0.2, 0.6\}$ in two of the four pairs, and so do the admissible $\{0.2, 0.6\}$ of s_B against $\{0.1, 0.8\}$. The coverage-score laws differ, $c_A \in \{0.1, 0.8\}$ against $c_B \in \{0.2, 0.6\}$. At $T = 1/2$ the population threshold is the smaller coverage score of each, 0.1 and 0.2 respectively, and both filters are valid at exactly risk $1/2$; but s_A keeps only the admissible candidate of the first question, $N_A = 1/2$, while the threshold of s_B also admits that question's inadmissible candidate at 0.1, giving $N_B = 1$. Equal AUC, twice the sets. The mechanism is that AUC records only which within-question comparison each score wins, while N also charges for where that comparison sits relative to the calibrated cut. This is the formal reason behind the ranking-not-AUC principle of Xi et al. (2025), which we restate and do not claim as ours.

A.3 The equivalence

Restatement (Theorem 1). *Assume (A0) and minimize $\mathbb{E}|S(X)|$ over all measurable, possibly randomized filters $S : X \mapsto 2^{[m]}$ subject to $\mathbb{P}(S \text{ covers no admissible answer}) \leq T$.* (i) *In the single-admission regime ((A2) unused) the size-optimal risk-T filter is Chow's rule: commit to $\arg\max_j q_j$ if $\max_j q_j \geq \theta_T$ and abstain otherwise, randomizing at the budget boundary. The optimal nonconformity score is $s_j = -q_j$, conformal calibration selects θ_T, and the achievable size-risk frontier is the risk-coverage curve of $p = \max_j q_j$.* (ii) *In the multi-admission regime, under (A2) the optimum is the greedy posterior filter, adding candidates in decreasing q_j while $q_j \prod_{k \in S}(1 - q_k) \geq 1/\lambda_T$; without (A2) it is $(1 - 1/e)$-approximate.*

Proof. Introduce the Lagrangian over possibly randomized filters,

$$\min_S \ \mathbb{E}\big[\, |S(X)| - \lambda \left(\operatorname{cov}(S(X)) - (1-T)\right) \big], \qquad \operatorname{cov}(S) := \mathbb{P}(\exists \text{ admissible in } S \mid \Phi).$$

The expectation separates over X, so it suffices to solve, per question, $\min_{S \subseteq [m]} |S| - \lambda \operatorname{cov}(S)$. Under (A2), $\operatorname{cov}(S) = 1 - \prod_{j \in S}(1 - q_j)$, and adding a candidate $j \notin S$ changes the objective by $1 - \lambda q_j \prod_{k \in S}(1 - q_k)$. This marginal cost increases as S grows (each factor $1 - q_k \leq 1$ shrinks

the product) and, among single additions, is smallest for the largest q_j. The optimal S therefore adds candidates in decreasing q_j and stops at the first j whose gain $\lambda q_j \prod_{k<j}(1-q_k)$ falls below 1. The dual risk $R(\lambda) = \mathbb{E}[1 - \mathrm{cov}(S_\lambda)]$ is monotone in λ but a step function, since per-question size is integer valued, so a generic T lies strictly between two attainable risk levels and the exact optimum is the Neyman–Pearson randomization between the two bracketing deterministic filters, which attains risk exactly T. If even the top candidate fails the test the optimal set is empty, which is abstention. Statement (i) is the restriction to $|S| \le 1$, where the factors $\prod_k(1-q_k)$ never appear and (A2) is unused, and is verbatim Chow's threshold rule. The greedy order is the order induced by $s_j = -q_j$; a conformal threshold on $-q$ is the level set $\{q_j \ge \theta\}$, and by Lemma 1 the conformal quantile selects the θ at which the empirical risk meets T, that is $\theta = 1/\lambda_T$. Without (A2), cov is a monotone submodular coverage function, so greedy attains $(1 - 1/e)$ of the optimal coverage at each size, the standard submodular guarantee, and Bayes-optimality is claimed only in regime (i). □

The reject-option equivalence itself is classical (Chow, 1970; Franc et al., 2023; Szabadváry et al., 2025); what is instantiated here is its form for multi-candidate answer-set filtering under the "at least one admissible" loss. Note that the optimum consumes only the *ranking* of q, so by Lemma 2 a better-calibrated verifier cannot change the sets unless it reorders q.

A.4 The floor, and the closed-form limit

Restatement (Corollary 1)**.** *Assume (A0) and $e_0 \le T$. Let the water-filling threshold p_τ solve $\int_{p \le p_\tau} p \, \mathrm{d}F_p(p) = (T - e_0)_+$ and let $G(T) = F_p(p_\tau)$. Any threshold filter with population validity T has $N \ge 1 - T$, and the minimum expected set size over $\mathcal{S}_1$, the filters returning at most one answer per question, is $N_{\mathrm{Bayes}}(T) = 1 - G(T) \in [1 - T,\, 1]$, equal to $1 - T$ exactly when the budget is spent entirely on zero-confidence mass, and then only for a filter free to randomize at an atom of c.*

Proof. The floor. On the event $\{c(X) \le t^\star\}$ the best admissible candidate has score at most $t^\star$, so it lies in $C_{t^\star}(X)$ and $|C_{t^\star}(X)| \ge 1$; since $\mathbb{P}(c \le t^\star) = 1 - T$ by definition of $t^\star$, $N(s) = \mathbb{E}|C_{t^\star}| \ge \mathbb{E}[\mathbf{1}\{c \le t^\star\}] = 1 - T$ for any score with population validity T. Approaching it: let $\rho(X)$ return one admissible candidate whenever one exists and let the oracle score give $\rho(X)$ the value -1 and every other candidate $+1$. Then $c(X) = -1$ exactly on $\{\exists \text{ admissible}\}$, and the kept set is $\{\rho(X)\}$ there and empty elsewhere. The optimum is therefore a single-representative-surfacing rule with abstention: a *second* correct paraphrase is pure waste, so the size-optimal rule is not a correctness classifier.

Why the floor needs randomization. That oracle score puts an atom of mass $\mathbb{P}(\exists \text{ admissible}) \ge 1 - T$ at -1, so it violates (A3), and a deterministic threshold cannot cut inside an atom: every $t \in [-1, 1)$ gives the same filter, of risk $1 - \mathbb{P}(\exists \text{ admissible}) \le T$ and size $N = \mathbb{P}(\exists \text{ admissible})$, which equals $1 - T$ only when (A0) binds with equality. To reach $1 - T$ in general the filter must abstain on part of that atom as well, that is randomize there (or split the atom by a tie-break). The gap between the oracle floor and what a deterministic conformal filter attains is exactly that randomization, and the same obstruction, seen from the other side, is Corollary 2.

The value. Work in $\mathcal{S}_1$. A filter there answers on a set K of questions and abstains off it, and by Theorem 1(i) it is optimal to answer with the top candidate, so its risk is $\mathbb{E}[\mathbf{1}_K(1-p)] + \mathbb{P}(K^c) = 1 - \mathbb{E}[\mathbf{1}_K\, p]$ and its size is $\mathbb{P}(K)$. Minimizing $\mathbb{P}(K)$ subject to $\mathbb{E}[\mathbf{1}_K\, p] \ge 1 - T$ is a continuous knapsack, linear in $\mathbf{1}_K$ in both objective and constraint, so its optimum fills by decreasing value density $1/p$: answer on the most-confident questions, abstain on $\{p \le p_\tau\}$ of mass $G(T) = F_p(p_\tau)$, randomizing at p_τ. Equivalently, start from the reference policy "keep the top candidate everywhere", of size 1 and risk e_0, and spend the slack $T - e_0$ abstaining, one unit of size for p units of risk. Either way $N_{\mathrm{Bayes}} = 1 - G(T)$, and $N_{\mathrm{Bayes}} \ge 1 - T$ because $p \le 1$ forces $\mathbb{P}(K) \ge \mathbb{E}[\mathbf{1}_K\, p] \ge 1 - T$.

Why the value is scoped to $\mathcal{S}_1$. Outside $\mathcal{S}_1$ enlarging one question's set lowers that question's risk and so frees budget to abstain on another, and the two moves can compose to a strictly smaller expected size than $1 - G(T)$ at the same risk. Two questions of equal mass, the first with posteriors $(0.01, 0)$ and the second with $(0.5, 0.5)$, give $e_0 = 0.745$ and $\mathbb{P}(\exists \text{ admissible}) = 0.38$, so (A0) holds at $T = 0.745$ and there is no atom at $p = 0$; the water-filling value is 1, yet abstaining on the first question and mixing, on the second, between keeping both candidates and keeping one meets the

same budget at expected size 0.52. We therefore state the value over $\mathcal{S}_1$, which is where the paper's claims are made and where (A2) is vacuous; Theorem 1 governs the unrestricted optimum instead.

The equality case. If $p \in \{0, 1\}$ almost surely then $e_0 = \mathbb{P}(p = 0)$; abstaining on the $p = 0$ questions costs no risk and removes their mass, and the remaining budget $T - e_0 \geq 0$ is spent on $p = 1$ mass, so $G(T) = T$ and $N_{\text{Bayes}} = 1 - T$. Any interior mass of p makes the value density $1/p$ finite there, so the budget buys less mass, $G(T) < T$ and $N_{\text{Bayes}} > 1 - T$. Both readings need the randomization above, since the budget is exhausted strictly inside the $p = 1$ atom. This is asserted only for $e_0 \leq T$: for $T < e_0$ the water-filling formula does not apply. □

Corollary 1 is the expected size of the Bayes-optimal level-set rule of Sadinle et al. (2019), whose optimality is already two-sided, re-expressed as a survival functional of the confidence statistic; Dhillon et al. (2024) is the closest expected-size precedent. No new optimality theorem is claimed.

A.5 THE REGIME LAW

Restatement (Theorem 2)**.** *Assume (A0) and $\mathbb{P}(p = 0) = 0$. Then $e_0 < T \iff \min_{\mathcal{S}_1} N = 1 - G(T) < 1$, $e_0 = T \iff \min_{\mathcal{S}_1} N = 1$, and $e_0 > T \iff \mathcal{S}_1$ has no valid member, in which case every valid filter keeps two or more candidates on a set of positive mass and, under (A2), the optimum is $\mathbb{E}_X[k_{\lambda_T}(X)]$, the expected greedy posterior keep-count at the risk-binding price.*

Proof. The three hypotheses $e_0 < T$, $e_0 = T$, $e_0 > T$ are exhaustive and mutually exclusive, and so are the three conclusions, since $\mathcal{S}_1$ either has a valid member with minimum size below one, or with minimum size one, or none at all. It therefore suffices to prove the three forward implications; the converses follow by trichotomy, and no separate argument for them is needed.

The one-answer floor on risk. Every $S \in \mathcal{S}_1$ answers on a set K and abstains off it, so its risk is $\mathbb{P}(K^c) + \mathbb{E}[\mathbf{1}_K(1 - q_{j(X)})] \geq \mathbb{E}[1 - p] = e_0$, with equality exactly for the reference policy "keep the top candidate everywhere", whose size is 1. So $\mathcal{S}_1$ contains a valid filter if and only if $e_0 \leq T$, and the smallest risk it can reach is e_0.

Case $e_0 > T$. By the display above every member of $\mathcal{S}_1$ has risk at least $e_0 > T$, so none is valid. Any valid filter must therefore keep two or more candidates on a set of positive mass, and by Theorem 1 the unrestricted optimum is the greedy posterior filter at the price λ_T making the risk bind, of size $\mathbb{E}_X[k_{\lambda_T}(X)]$.

Case $e_0 < T$. The reference policy is strictly feasible, and by Corollary 1 the minimum over $\mathcal{S}_1$ is $1 - G(T)$ with $\int_{p \leq p_\tau} p \, \mathrm{d}F_p = T - e_0$. Since $p \leq 1$, the abstained mass satisfies $G(T) = F_p(p_\tau) \geq \int_{p \leq p_\tau} p \, \mathrm{d}F_p = T - e_0 > 0$, so $\min_{\mathcal{S}_1} N \leq 1 - (T - e_0) < 1$.

Case $e_0 = T$. The slack vanishes, so $\int_{p \leq p_\tau} p \, \mathrm{d}F_p = 0$. Under $\mathbb{P}(p = 0) = 0$ the integrand is strictly positive F_p-almost everywhere, so the abstained set carries no mass: $G(T) = 0$ and $\min_{\mathcal{S}_1} N = 1$. □

The zero-atom hypothesis is not vacuous. Drop $\mathbb{P}(p = 0) = 0$ and the boundary case fails, inside the paper's own assumption set. Put mass T at $p = 0$, the questions no filter can ever cover, and mass $1 - T$ at $p = 1$. Then $e_0 = T$, but abstaining on the zero-confidence mass raises its risk contribution from 1 to 1 and so costs nothing, giving $G(T) = T$ and $\min_{\mathcal{S}_1} N = 1 - T < 1$: the equivalence $\min_{\mathcal{S}_1} N = 1 \iff e_0 = T$ is contradicted. This configuration has $\mathbb{P}(\exists \text{ admissible}) = 1 - T$, so it satisfies (A0) with equality, and the method encodes exactly this mass as $c = +\infty$ (Equation (1)). The hypothesis says that abstention is nowhere free, equivalently that c has no atom at $+\infty$; the alternative reading, stating the law on the coverable sub-population, is equivalent and we do not use it.

Caveats, transcribed. The exact form $\mathbb{E}[k_{\lambda_T}]$ in the ranking regime uses (A2); without it, by submodularity, N_{Bayes} lies within $(1 - 1/e)$ of it. A *scalar* closed form for the ranking-regime value, analogous to $1 - G(T)$, is not established here and is **[OPEN-1]** of Section 3.4; what is established is the variational characterization.

A.6 SCALE-FREE SCORES, AND THE ATOM THAT BLOCKS ABSTENTION

Restatement (Corollary 2)**.** *Call a score* scale-free *if it is a fixed increasing function of the within-question rank alone. Fix $T < 1$, a deterministic threshold, and a common candidate count m across questions. Then $c(X)$ lives on the m-point rank grid and has an atom at its bottom of mass $\pi_0 = \mathbb{P}$(top-ranked candidate admissible); $|C_t(X)|$ is the same integer for every X; and every valid such filter has $N \in \{1, \dots, m\}$, the size-minimizing valid choice giving $N = 1$ at the constant realized risk $1 - \pi_0$ for every $T \geq 1 - \pi_0$. That is above $N_{\mathrm{Bayes}}(T)$ at every risk level when $e_0 < T$, and unconstrained when $e_0 > T$.*

Proof. Write $\rho_j(X) \in \{0, \frac{1}{m-1}, \dots, 1\}$ for the within-question normalized rank of candidate j, so that $s_j = g(\rho_j)$ for a fixed increasing g and, every question carrying the same candidate count m, the value set $\mathcal{G} = g(\{0, \frac{1}{m-1}, \dots, 1\})$ is the *same* in every question. This is the whole content of "no cross-question scale": a score value carries the rank it encodes and nothing about how confident that rank is.

(i) The atom. $c(X) = g(\rho_{j^\star}(X))$ where $j^\star$ is the best-ranked admissible candidate, so c takes values in $\mathcal{G} \cup \{+\infty\}$ and $\mathbb{P}(c = g(0)) = \mathbb{P}(A = 1 \text{ at rank } 1) = \pi_0 > 0$. The coverage-score law is discrete with an atom at its bottom, which is exactly the situation (A3) excludes.

(ii) The count is constant. For any t, $|C_t(X)| = \#\{j : g(\rho_j(X)) \leq t\} = \#\{v \in \mathcal{G} : v \leq t\}$, which does not depend on X. The filter returns the same number of answers on every question, so it cannot answer on some and abstain on others: abstention is all or nothing. This is the one step that uses the common candidate count, and the paragraph after the proof records what survives without it.

(iii) The floor at one. If $t < g(0)$ then $C_t(X) = \emptyset$ for every X and the risk is $1 > T$, so validity forces $t \geq g(0)$ and hence $N = |C_t| \geq 1$, while trivially $N \leq m$. Now $t = g(0)$ has risk $\mathbb{P}(c > g(0)) = 1 - \pi_0$, so it is valid whenever $T \geq 1 - \pi_0$, and it is then the smallest valid threshold, hence the one a size-minimizing conformal rule selects. There $N = 1$ and the realized risk is $1 - \pi_0$, *neither depending on* T: the filter does not respond to the risk budget at all. Finally, by Theorem 2, $e_0 < T$ gives $\min_{\mathcal{S}_1} N = 1 - G(T) < 1 \leq N$, so the scale-free filter is strictly above the one-answer optimum at every risk level in the abstention regime; and $e_0 > T$ makes $\mathcal{S}_1$ empty, so a floor at one candidate costs nothing there. □

Unequal candidate counts. Step (ii) is the only place the common candidate count enters, and it is the only claim that needs it. If the count varies with the question, write $m'(X)$ for the count on question X; the rank grid varies with it, so $|C_t(X)|$ is no longer a single integer, and N is an average over questions rather than that integer. Everything the corollary rests on survives. The top-ranked candidate of every question still sits at the common bottom value $g(0)$, so c still has an atom there of mass π_0; any $t < g(0)$ still returns the empty set on every question, at risk $1 > T$, so validity still forces $t \geq g(0)$; and $t = g(0)$ still keeps exactly the top-ranked candidate of each question, so it is valid whenever $T \geq 1 - \pi_0$ and has $N = 1$ at the constant realized risk $1 - \pi_0$. Hence $N \geq 1$ with equality at the smallest valid threshold, which is the floor the corollary is about and the only part Section 5 measures. The deployment of Section B draws a fixed $m = 20$ candidates per question and scores an absent candidate at $+\infty$, so its counts are equal by construction up to those absences, and the stronger constant form applies there as well.

What this does and does not say. The mechanism is the atom in the coverage-score law, not ranking quality, and the two must not be conflated. Nothing above says that ranking correctness better makes a filter worse: by Theorem 1 the size-optimal score is $s = -q$, the perfect correctness *posterior*, which carries a cross-question scale and attains $1 - G(T)$. What the corollary says is that discarding that scale, keeping only the order inside each question, forfeits the entire abstention gain $G(T)$, and that this is fatal exactly where the optimum lies below one answer and irrelevant exactly where it lies above. Restoring abstention requires cutting inside the atom, which a deterministic threshold cannot do: either the threshold randomizes there, as in the achievability half of Corollary 1, or the atom is split by a tie-break. With normalized ranks the bands sit $1/(m-1)$ apart, so an additive tie-break of magnitude $\varepsilon < 1/(m-1)$ splits the bottom atom into an interval without moving any candidate across a rank boundary, and the filter becomes a {abstain, single answer} selective predictor; Section B records the deployed constant.

Relation to Xi et al. (2025): a reversal, not a sharpening. Their result is that higher confidence yields smaller prediction sets in expectation. The statement above runs in the opposite direction in the same variable, and the reason is a difference of setting rather than a correction: their prediction sets are never empty, so confidence can only shrink a set that must contain at least one label, whereas here the empty set is admissible and is the cheapest set of all, which is what makes the inability to abstain expensive. We therefore label the relation a contrast under abstention and not a sharpening of their theorem. This is the content the method exploits, in the gate of Equation (7).

A.7 Finite-sample versions

Fixed score. Lemma 1 is already finite-sample. The efficiency optimum shifts accordingly: $\hat{t}$ targets the $(1 - T + O(1/n))$-quantile of c, so the optimum is $1 - T + O(1/n)$ and every population statement above holds up to that shift.

Learned score, and the exchangeability it breaks. If a score is fit on data, the calibration and test coverage scores are no longer exchangeable and Lemma 1 does not apply. Validity is restored either by split conformal calibration, fitting the score on a training fold disjoint from calibration and test, so that conditionally on that fold the score is a fixed function and Lemma 1 applies for almost every training fold, or by cross-conformal and CV+ variants at the standard inflation. ACF-LV takes the first route, which is why its verifier is cross-fit.

Operating-point selection, and the guarantee it buys. When the operating point is *selected* from a finite family, each candidate is treated as a hypothesis "risk $> T$" and certified. With calibration size n and empirical risk $\hat{R}$, the Hoeffding–Bentkus p-value

$$P_{\mathrm{HB}} = \min\Big(\exp\big(-n\,\mathrm{KL}(\min(\hat{R}, T)\,\|\,T)\big),\ e \cdot \mathrm{Binom}(\lceil n\hat{R}\rceil;\, n,\, T)\Big)$$

controls the selected configuration by Bonferroni over a grid of size G: choosing any configuration with $P_{\mathrm{HB}} \le \delta/G$ guarantees $\mathbb{P}(\text{selected configuration has true risk} > T) \le \delta$ (Learn-then-Test, Bates et al., 2021; Angelopoulos et al., 2026). This converts the population threshold into a PAC-valid finite-sample certificate, and it is what makes single-stage end-to-end selection honest. The guarantee *type* differs from the marginal default: a δ is stated, and it is sourced from the Bonferroni union over the grid.

B Implementation and hyperparameters

This section gives the implementation details used by the experiments in Section 5.

Algorithm 1 The ACF-LV procedure.

Require: training data when the verifier is used, calibration labels A_j, candidate features Φ_j, entropies u, risk level T, margin grid $\mathcal{K}$

1: **if** $\mathrm{AUC}(-u, \mathbf{1}[A = 1]) < 1/2$ on calibration labels **then**
2: fit the verifier on the training fold and set s to within-question rank of its score
3: **else**
4: set s to the entropy rank with the cross-question tie-break of Equation (6)
5: **end if**
6: compute $c_i = \min\{s_j \mid A_j = 1\}$ for each calibration question, with $+\infty$ when none is admissible
7: for ACF, choose the smallest margin $\kappa \in \mathcal{K}$ whose held-out validation risk satisfies the target and recompute $\hat{t}$ from Equation (5)
8: for ACF-LTT, certify the selected threshold over the same grid with Learn-then-Test
9: return $C_{\hat{t}}(X) = \{j \mid s_j \le \hat{t}\}$ at test time

Sampling and admissibility. Each question produces at most $m = 20$ candidates by multinomial sampling at temperature 1.0, following the published defaults of Wang et al. (2026). The default admissibility rule is reference similarity of at least 0.6. The sampling level is $\alpha = 0.05$ on ScienceQA and TriviaQA and $\alpha = 0.25$ on CoQA. We sweep $\beta \in \{0.05, 0.10, \ldots, 0.50\}$ and use $T = \alpha + \beta - \alpha\beta$ as the composed risk level.

Unified splits. The primary experiment uses one three-way split for every method. The training fraction is 0.3, the calibration fraction is 0.2, and the test fraction is 0.5. We use 100 reshuffles with identical folds across all arms. The training fold is only consumed by verifier-based arms, but it is reserved for every method so that set sizes are directly comparable. Earlier diagnostic experiments used a 0.5 calibration split when no verifier was trained. Those results are retained below for continuity and robustness checks, but they are not mixed with the unified headline comparison.

Entropy score. The score in Equation (6) uses the within-question normalized rank of length-normalized predictive token entropy. The rank lies in $[0, 1]$ with spacing $1/(m-1)$. The quantity $z_j \in [0, 1]$ is a min-max normalization of entropy across the evaluation cell. We set $\epsilon = 0.05$. The sweep in Section D shows that every tested value at or below the structural rank gap gives effectively identical results.

Validity margin. The candidate margins are $\mathcal{K} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0\}$. For the marginal variant, the calibration data are split into a fit fold and a validation fold. The smallest margin whose validation empirical risk meets the target is selected, after which the threshold is recomputed on the full calibration set. For the high-probability variant, the same finite grid is certified with Hoeffding–Bentkus p-values and Bonferroni correction as described in Section A.7.

Verifier. The verifier is a gradient-boosted LambdaRank model (Ke et al., 2017) with 200 trees, 15 leaves, learning rate 0.05, and a minimum of 20 samples per leaf. It uses eight reference-free features for each candidate. These are the mean, standard deviation, maximum, minimum, and token count of the per-token entropy sequence, the generation order, the normalized frequency of the candidate surface form among the sampled answers, and an indicator for the modal answer. The verifier is fitted on a training fold that is disjoint from calibration and test. Its output is converted to within-question rank and therefore carries no cross-question tie-break.

Datasets and backbones. The 15-cell grid crosses ScienceQA (Lu et al., 2022), TriviaQA (Joshi et al., 2017), and CoQA (Reddy et al., 2019) with Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-14B-Instruct (Qwen et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and OpenChat-3.5 (Wang et al., 2024). Candidate generation is the only accelerator-heavy step. Calibration, filtering, and tree fitting are lightweight once candidates and entropy statistics have been cached.

C Estimator validation and legacy grid

The unified estimator was checked against values reported by the earlier implementation before it was used for the new comparisons.

Table 5: Estimator validation. Representative cells reproduce the previously reported set sizes exactly.

Cell	Arm	Unified implementation	Reported	Difference
CoQA × Qwen2.5-7B	SAFER	9.985	9.985	0.0000
CoQA × Qwen2.5-7B	ACF-LTT	3.179	3.179	0.0000
TriviaQA × OpenChat-3.5	SAFER	7.296	7.296	0.0000
TriviaQA × OpenChat-3.5	ACF-LTT	1.302	1.302	0.0000
SciQA × Llama-3.1-8B	SAFER	1.440	1.440	0.0000
SciQA × Llama-3.1-8B	ACF-LTT	0.876	0.876	0.0000

The original 0.5 calibration experiment remains useful because it covers every dataset-model cell with the larger calibration fold. It should be read as an entropy-only ACF-LTT comparison rather than as the end-to-end ACF-LV result.

The reproduced two-stage results were checked against the test-time error rates published by Wang et al. (2026). Across all 15 cells and eight overlapping risk levels, all 120 comparisons fall within the reported standard deviation and the maximum absolute discrepancy is 0.0118.

Table 6: Original 15-cell comparison at the 0.5 calibration split. The complete verifier-gated method is evaluated under the unified protocol in Table 1.

		SAFER		ACF-LTT			ACF
Dataset	Backbone	N	val.	N	val.	ΔN	N
SciQA	Qwen2.5-7B	.739	82%	.794	100%	+7%	.730
	Llama-3.1-8B	1.440	76%	.876	100%	-39%	.722
	Qwen2.5-3B	1.386	77%	.815	99%	-41%	.731
	Qwen2.5-14B	.729	82%	.793	100%	+9%	.728
	OpenChat-3.5	1.296	79%	.808	100%	-38%	.737
TriviaQA	Qwen2.5-7B	3.291	77%	.950	100%	-71%	.763
	Llama-3.1-8B	3.180	81%	1.039	100%	-67%	.785
	Qwen2.5-3B	5.003	73%	1.426	100%	-71%	.893
	Qwen2.5-14B	1.983	81%	.892	100%	-55%	.760
	OpenChat-3.5	7.296	68%	1.302	100%	-82%	.863
CoQA	Qwen2.5-7B	9.985	80%	3.179	100%	-68%	2.683
	Llama-3.1-8B	1.186	99%	.854	100%	-28%	.678
	Qwen2.5-3B	7.392	84%	2.889	100%	-61%	2.858
	Qwen2.5-14B	10.292	71%	3.249	100%	-68%	2.612
	OpenChat-3.5	1.385	99%	.880	100%	-36%	1.001

Table 7: Per-cell validity ledger for the original 0.5 calibration analysis. The final column gives the largest ratio of mean test error to its risk bound across the ten risk levels.

Dataset	Backbone	n	SAFER	ACF	ACF-LTT	worst mean-EER/bound
SciQA	Qwen2.5-7B	789	82%	64%	100%	.976
	Llama-3.1-8B	975	76%	59%	100%	.985
	Qwen2.5-3B	731	77%	63%	99%	.984
	Qwen2.5-14B	846	82%	65%	100%	.977
	OpenChat-3.5	972	79%	66%	100%	.981
TriviaQA	Qwen2.5-7B	1592	77%	67%	100%	.977
	Llama-3.1-8B	1919	81%	65%	100%	.980
	Qwen2.5-3B	1464	73%	58%	100%	.991
	Qwen2.5-14B	1742	81%	65%	100%	.985
	OpenChat-3.5	1932	68%	59%	100%	.994
CoQA	Qwen2.5-7B	1872	80%	70%	100%	.987
	Llama-3.1-8B	1960	99%	66%	100%	.982
	Qwen2.5-3B	1831	84%	63%	100%	.985
	Qwen2.5-14B	1932	71%	57%	100%	.992
	OpenChat-3.5	1955	99%	63%	100%	.986
mean over cells / max in final column			80.5%	63.4%	99.8%	.994

D SENSITIVITY TO THE TIE-BREAK CONSTANT

Corollary 2 predicts a structural change when the cross-question tie-break becomes large enough to alter within-question rank order. With $m = 20$, the rank gap is $1/(m-1) = 0.0526$. We therefore sweep values on both sides of this point.

Table 8: Sensitivity to ϵ. Results are effectively invariant below the structural rank gap.

ϵ	Mean set size	Per-split validity	Max cell deviation from 0.05
0.01	1.383	99.8%	0.000064
0.02	1.383	99.8%	0.000045
0.03	1.383	99.8%	0.000033
0.05	1.383	99.8%	0.000000
0.0526	1.383	99.8%	0.000009
0.06	1.383	99.8%	0.000656
0.10	1.386	99.8%	0.035378
0.20	1.390	99.8%	0.059417

The largest per-cell deviation from the deployed value is 6.4×10^{-5} over the tested range from 0.01 through 0.0526. All 15 cells change once the sweep moves above the structural gap. This supports treating $\epsilon = 0.05$ as a structural choice rather than a tuned hyperparameter.

E RISK-LEVEL DEPENDENCE

The main paper reports averages over ten risk levels. Here we summarize the paired comparison at each level using the same 100 reshuffles for both methods. The method is smaller at every risk level on 9 of the 11 pairable cells. Of the 110 paired cell-risk points, 109 are significant at the 99% level over the 100 reshuffles. The two exceptions are ScienceQA cells where both procedures already return fewer than one answer. TriviaQA with OpenChat-3.5 and the three CoQA Qwen cells are omitted because the baseline is infeasible at some risk levels.

Table 9: Paired risk-level comparison. Changes are method minus baseline set size.

Dataset	Backbone	Smaller levels	Largest change	Smallest change
SciQA	Qwen2.5-7B	0/10	+0.03	+0.08
SciQA	Llama-3.1-8B	10/10	-0.83	-0.25
SciQA	Qwen2.5-3B	10/10	-0.89	-0.27
SciQA	Qwen2.5-14B	0/10	+0.04	+0.08
SciQA	OpenChat-3.5	10/10	-0.78	-0.25
TriviaQA	Qwen2.5-7B	10/10	-3.66	-1.19
TriviaQA	Llama-3.1-8B	10/10	-2.69	-1.52
TriviaQA	Qwen2.5-3B	10/10	-5.89	-1.57
TriviaQA	Qwen2.5-14B	10/10	-1.65	-0.61
CoQA	Llama-3.1-8B	10/10	-0.55	-0.04
CoQA	OpenChat-3.5	10/10	-0.77	-0.16

F CONFIDENCE LEVEL

We repeat the high-probability calibration at delta 0.05 to match the confidence parameter used in the first stage of the two-stage reference procedure.

Table 10: Sensitivity to delta. Halving delta changes mean set size only modestly while maintaining high observed per-split validity and certification.

Delta	Mean set size	Per-split validity	Certification rate
0.05	1.421	99.9%	99.9%
0.10	1.383	99.8%	100.0%

Moving from delta 0.1 to delta 0.05 increases mean set size from 1.383 to 1.421, a change of 2.7%. Observed per-split validity is 99.9% and the certification rate is also 99.9% at delta 0.05.

G GATE SPECIFICITY

The calibration gate does not fire on three TriviaQA cells whose entropy AUC is above 0.5. To test whether the gate misses a useful verifier, we fit the verifier on these cells anyway.

Table 11: Forced-verifier test on cells rejected by the gate. The verifier increases set size in all three cases.

Dataset	Backbone	Entropy AUC	Rank score	Forced verifier	Change
TriviaQA	Llama-3.1-8B	.820	1.268	1.304	+2.9%
TriviaQA	Qwen2.5-3B	.552	1.978	2.094	+5.9%
TriviaQA	OpenChat-3.5	.543	1.873	1.881	+0.4%

The forced verifier increases expected set size on all three cells. We therefore observe no false negative of the gate on this grid in the operational sense of declining a verifier that would have reduced set size.

H SAMPLING-BUDGET TRADE-OFF

Both procedures are evaluated on the same cached pool of 20 candidates, so the raw generation work in the experiments is identical by construction. A deployed system would not necessarily draw all 20. The calibrated sampling budget therefore gives the more relevant comparison.

The proposed method uses a larger budget on 9 of 15 cells. Mean budgets are nevertheless close at 6.38 for the proposed method and 6.26 for the reference procedure. The distribution is highly nonuniform. On the three CoQA Qwen cells, the reference procedure uses between 12.68 and 15.99 candidates while our method uses between 3.01 and 3.37. On several ScienceQA cells the direction reverses. The set-size improvement should therefore not be interpreted as a uniform reduction in generation cost.

I CORRECTNESS CRITERIA AND CALIBRATION SIZE

The default admissibility rule is only one operational definition of correctness. We therefore repeat the analysis with the four criteria used by Wang et al. (2026).

Across the 49 feasible cells, the strict comparison favors ACF-LTT on 35. The count rises to 45 when a two percentage point tolerance is used. Median set-size reductions range from 21% to 55% across the four criteria.

We also vary the calibration fraction from 0.5 to 0.1 while keeping the rest of the pipeline fixed.

At the smallest calibration fraction, ACF-LTT remains smaller on 14 of 15 cells and observed per-split validity remains 100.0%. The price is a larger answer set and a higher rate of uncertified configurations, which is consistent with the stronger finite-sample penalty at small calibration sizes.

J EXTENDED RELATED WORK

Guarantees beyond exchangeability. Split conformal prediction provides the calibration construction used here (Papadopoulos et al., 2002; Lei et al., 2018; Vovk et al., 2022). Conformal risk control extends the same logic to monotone losses (Angelopoulos et al., 2024). These guarantees are marginal and assume exchangeability. Conditional validity and distribution shift require different tools (Vovk, 2012; Tibshirani et al., 2019; Barber et al., 2023). Selection by conformal p-values (Jin & Candes, 2023) is also close in form because it can return an empty selection, though it controls a different error criterion.

Uncertainty in language generation. Token entropy is a sequence-level uncertainty signal rather than a direct statement about semantic correctness. Semantic uncertainty aggregates equivalent generations before measuring uncertainty (Kuhn et al., 2023; Farquhar et al., 2024). Sampling-based consistency methods provide a related signal without requiring logits (Manakul et al., 2023). Our

theory applies to any score. A semantic score that retains a cross-question confidence scale would therefore be a natural alternative instantiation.

Operational correctness. The admissibility label used for calibration is a modeling choice. Long-form factuality can be evaluated through atomic-fact decomposition (Min et al., 2023), while model-based judging is common but carries known biases (Zheng et al., 2023). We treat the correctness-criterion experiment as a sensitivity analysis for this reason rather than as four independent confirmations of the same ground truth.

Table 12: Calibrated sampling budgets. The ratio is the proposed budget divided by the two-stage budget. Returned set sizes from the original 0.5 calibration analysis are included to show the relation between generation budget and output size.

Dataset	Backbone	Base budget	ACF budget	Ratio	Base size	ACF size
SciQA	Qwen2.5-7B	1.01	2.50	2.48	.739	.794
SciQA	Llama-3.1-8B	2.31	9.46	4.10	1.440	.876
SciQA	Qwen2.5-3B	2.09	6.96	3.33	1.386	.815
SciQA	Qwen2.5-14B	1.00	2.95	2.95	.729	.793
SciQA	OpenChat-3.5	1.96	7.09	3.62	1.296	.808
TriviaQA	Qwen2.5-7B	5.70	5.55	.97	3.291	.950
TriviaQA	Llama-3.1-8B	5.24	10.46	2.00	3.180	1.039
TriviaQA	Qwen2.5-3B	10.99	13.53	1.23	5.003	1.426
TriviaQA	Qwen2.5-14B	3.06	3.33	1.09	1.983	.892
TriviaQA	OpenChat-3.5	12.64	7.21	.57	7.296	1.302
CoQA	Qwen2.5-7B	15.26	3.35	.22	9.985	3.179
CoQA	Llama-3.1-8B	2.00	15.87	7.94	1.186	.854
CoQA	Qwen2.5-3B	12.68	3.01	.24	7.392	2.889
CoQA	Qwen2.5-14B	15.99	3.37	.21	10.292	3.249
CoQA	OpenChat-3.5	2.00	1.12	.56	1.385	.880

Table 13: Sensitivity to the correctness criterion. Strict counts require smaller sets and observed per-split validity at least as high as the two-stage procedure. The tolerance column allows a two percentage point validity difference.

Criterion	threshold	α	strict	$\pm$2pt	median reduction
similarity	.6	.05	13/15	13/15	55%
Rouge-L	.3	.5	7/11	10/11	25%
entailment	.5	.5	7/12	12/12	21%
LLM-judge	.5	.5	8/11	10/11	31%

Table 14: Sensitivity to calibration size. The advantage narrows with less calibration data but does not reverse.

split	$n_{\text{cal}} \geq$	SAFER	ACF-LTT	ratio	smaller	no certificate
.5	365	3.77	1.38	.37×	13/15	0.0%
.4	292	3.84	1.47	.38×	13/15	0.3%
.3	219	4.08	1.61	.39×	15/15	0.6%
.2	146	4.27	1.90	.45×	15/15	1.6%
.1	73	4.72	3.05	.65×	14/15	6.5%