RISK CONTROL FOR COMPLETE EVIDENCE ACTIONS: UNIFORM AUTHORIZATION AND RETRIEVAL LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) systems must decide both which evidence to retain and whether to answer. Preserving each relevant snippet with high probability does not ensure that an answer retains all required support. RC-POBP uses one calibrated envelope to authorize all 2^K retained-subset proposals, including data-dependent choices. Given item scores, authorization requires an $O(K)$ scan. We characterize when minimal harmful sets permit this reduction. With incomplete retrieval, the envelope separates missing from deleted support and gives a sharp coverage ceiling for every gate restricted to the retrieved candidates. Experiments on HotpotQA, 2Wiki, and MuSiQue distinguish calibrated support preservation, learned evidence compression, and selective answer quality. On MuSiQue, the authorizer removes 47.40% of evidence with a 4.96% support-omission upper bound while improving answer quality over full evidence. In a frozen 2Wiki audit, a linked-retrieval policy meets the 5% support-risk target with an answer-error upper bound of 10.12%. With joint 95% confidence, its coverage is at least 43.04%, above the 37.11% ceiling for every gate on the original BM25 candidates satisfying the same support-risk target. Its correct-answer yield also exceeds this ceiling. The comparison identifies a retrieval bottleneck that downstream gate tuning alone cannot overcome.

1 INTRODUCTION

Retrieval-augmented generation (RAG) couples a language model with external evidence (Lewis et al., 2020). A deployed system may filter passages, stop adaptively, and answer from the resulting context. For a multi-hop question, retaining a passage that names an entity is insufficient if the passage linking that entity to the requested attribute is removed. Thus, certifying each passage separately does not certify the composed context: m marginal harm events bounded by α can have union probability $\min\{1, m\alpha\}$. The object requiring a guarantee is the complete action that deployment executes.

This paper studies two settings that share this problem but require different representations. In a fixed frame, all candidate items are known and a policy may choose any retained subset. In open retrieval, the candidate set itself varies with the query; support absent from the retrieved set cannot be restored by a downstream gate. The central questions are therefore: when can one calibration cover every complete action without enumerating the power set, and what coverage remains possible when retrieval is incomplete?

Our answer is a calibrated envelope over registered harmful actions. For fixed-frame support omission, singleton witnesses make the envelope computable by one linear scan, even when the proposed subset is chosen after observing scores. For general losses, the minimal harmful sets determine the exact envelope and its oracle-query complexity. For open retrieval, the same accounting separates missing support from support deleted after retrieval and yields a sharp coverage ceiling. These are representation and information limits; calibration supplies finite-sample risk control once the relevant event is fixed.

The experiments test support preservation on fixed frames, separate scoring from calibration in a matched comparison, and compare linked retrieval with the coverage ceiling for every downstream gate on the original BM25 candidates, including an oracle.

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.

The fixed-frame result concerns a prospectively registered item pool. The open-retrieval result assumes i.i.d. complete query–corpus traces, with arbitrary dependence within each trace. Losses are zero on abstention. Each population claim uses the stated question-sampling design and deployment distribution.

2 Related Work

Adaptive and multi-hop RAG methods change retrieved evidence, routing, or stopping (Asai et al., 2020; Xiong et al., 2021; Trivedi et al., 2023; Gutiérrez et al., 2025; Park et al., 2025; Li et al., 2026). Recent work further studies retrieval timing during reasoning, confidence- or sufficiency-guided evidence selection, and budget-aware retrieval and stopping (Guo et al., 2026; Qiu et al., 2026; Qian et al., 2026; Devarakonda et al., 2026; Kamthan, 2026; Kieback et al., 2026). These works focus on retrieval and stopping policies; our question is complementary: once a family of complete actions is specified, what event must be calibrated to cover it?

Risk-control methods calibrate registered losses: TRAQ and C-RAG address answer-set coverage and generation risk, while learn-then-test and conformal risk control apply to supplied losses under stated assumptions (Li et al., 2024; Kang et al., 2024; Angelopoulos et al., 2025; 2024). Recent work also studies counterfactual RAG-risk control, cascaded-RAG calibration, inference-time factuality control, trajectory-level agent risk, and compute-budgeted reasoning (Chen et al., 2024; Jia et al., 2026; Wang et al., 2026a; Opoku & Banahene, 2026; Wang et al., 2026b). These approaches address different specified losses and decision units; we characterize the event required to certify an entire registered action family.

Evidence-selection methods use different audit objects: Conformal-Embedding controls marginal snippet retention, context-cutoff methods stop processing after latent sufficiency signals, TASR selects ranked prefixes using empirical stability and correctness, and AC-RAC models action-conditional utility over a finite space (Chakraborty et al., 2026; Xie et al., 2025; Kieback et al., 2026; Zhu et al., 2026). Our family may contain every retained subset; when minimal harmful witnesses are singletons, its support-omission envelope has an implicit linear-time representation. A matched scorer-by-calibrator experiment separates calibration from score quality.

3 Problem Setup: Complete Actions and Certificate-Preserving Reductions

Both settings define harm on the complete action executed, including every decision affecting the returned answer.

Let x denote one deployment input and let Π be a registered family of proposal rules. Rule π returns a *complete action* $a_\pi(x)$: all evidence edits, routing choices, stopping decisions, generation, and fallback behavior that deployment executes together. Let $h_x(a) \in \{0,1\}$ be its task harm and $s_x(a) \in \mathbb{R}$ its admission score. At threshold γ, the harmful accepted event over the entire family is

$$E_{\Pi,\gamma}(x) = \max_{\pi\in\Pi} \mathbb{I}[h_x(a_\pi(x)) = 1,\ s_x(a_\pi(x)) > \gamma]\,. \tag{1}$$

Define the complete-action envelope

$$W_\Pi(x) = \sup\{s_x(a_\pi(x)) : \pi \in \Pi,\ h_x(a_\pi(x)) = 1\}, \qquad \sup\varnothing = -\infty. \tag{2}$$

Then $E_{\Pi,\gamma}(x) = \mathbb{I}[W_\Pi(x) > \gamma]$ for every x and γ. This identity is certificate-preserving: any upper-tail certificate for W_Π applies simultaneously to every registered proposal and to any data-dependent choice among them. It is also minimal. If a scalar $V(x)$ satisfies $E_{\Pi,\gamma}(x) \le \mathbb{I}[V(x) > \gamma]$ for every threshold, then $V(x) \ge W_\Pi(x)$; equality for every threshold forces $V(x) = W_\Pi(x)$. The remaining challenge is to compute this envelope without enumerating the action space.

For a question q, $\mathcal{C}_{q,r}$ contains the actions registered for source r. With annotated support S_q, our context-cutoff loss is $h_q^{\mathrm{stop}}(r,t) = \mathbb{I}[S_q \nsubseteq E_{r,t}]$. The complete executed action includes fallback and any post-admission transformation; its harm must remain inside the calibrated event.

Open retrieval has no common realized candidate set. Let $T = G(Q, \mathcal{W}, U)$ be the almost-surely finite trace generated from query Q, corpus $\mathcal{W}$, and all internal retrieval randomness U. Its length

and action identities may vary with the question, and its steps may be adaptively dependent. A frozen measurable policy π returns one complete action in T or the safe action $\perp$ (abstain). Before auditing, fix binary losses $\ell_j(Q, T, a)$ with $\ell_j(Q, T, \perp) = 0$. The annotation may encode incomplete support, factual error, policy violation, or human adjudication; it is revealed only after $\pi(T)$ is fixed. For one loss, define

$$R_{\mathrm{sel}}(\pi) = \Pr\{\ell(Q, T, \pi(T)) = 1 \mid \pi(T) \neq \perp\}, \qquad R_{\mathrm{end}}(\pi) = \Pr\{\ell(Q, T, \pi(T)) = 1\}. \quad (3)$$

Set $R_{\mathrm{sel}}(\pi) = 0$ when the population answer probability is zero. For a finite policy and loss library, the audit retains

$$\Psi_{\Pi,J}(Q, T) = \Big(\mathbb{I}[\pi(T) \neq \perp],\ \big(\ell_j(Q, T, \pi(T))\big)_{j=1}^{J}\Big)_{\pi\in\Pi}. \quad (4)$$

The complete trace is the sampling unit; $\Psi_{\Pi,J}$ retains its answer and loss coordinates. Summing these coordinates gives the counts used by Corollary 2. Per-trace reconstruction supports reuse across all registered decision problems; the mean-risk intervals themselves need only those aggregate counts.

4 Method: Complete-Action Authorization

With a fixed item frame, the method reduces the power-set event to a support maximum and restores items with scores at or below the calibrated threshold. With open retrieval, it freezes a complete answer-or-abstain policy and audits its losses on full query–corpus traces.

4.1 Closed-action envelope and calibration

A scorer $s_\theta(c) = s_\theta(q, r, c)$ ranks registered complete actions. It is fitted only on proper-training questions and their harm labels. Calibration harm labels are withheld from fitting and used only to form the maxima below and choose γ. For each calibration question and registered family $\mathcal{R}$, Eq. 2 specializes to

$$M(q) = \max\{s_\theta(c) : r \in \mathcal{R},\ c \in \mathcal{C}_{q,r},\ h_q(c) = 1\}, \quad (5)$$

with $M(q) = -\infty$ if no harmful registered action exists, and let $B_\gamma(q) = \mathbb{I}[M(q) > \gamma]$. This event records whether the registered family contains any harmful candidate above the deployed threshold. Write the ascending calibration order statistics as $M_{(1)} \leq \cdots \leq M_{(n)}$. For target risk α and calibration failure probability δ, choose the smallest index k satisfying

$$\Pr\{\mathrm{Binomial}(n, 1 - \alpha) \geq k\} \leq \delta, \quad (6)$$

and set $\gamma = M_{(k)}$, or $+\infty$ if no rank qualifies. Deployment admits only strict exceedances; fixed arbitration selects complete actions and never composes separately calibrated edits. Preregistered mutually exclusive strata use $\sum_g \delta_g \leq \delta$ and separate γ_g; a shared real-valued scorer and scale permit the conservative $\max_g \gamma_g$, whereas other scorers retain their stratum-specific thresholds. Actions that can co-occur must instead share $M(q)$ or a separate allocation of α. Score ranking affects coverage, not validity.

Proposal-agnostic power-set authorization. Let $C_q = \{1, \ldots, K_q\}$ be a registered item frame, $S_q \subseteq C_q$ its supporting items, and let $s_q(i)$ be an item deletion score, with larger values favoring deletion. Calibrate the question-level score

$$Z(q) = \max_{i \in S_q} s_q(i), \quad (7)$$

using the same order-statistic rule. A proposal $P_q \subseteq C_q$ may be any retained subset, including the output of an adaptive or data-dependent policy. Its authorized output is

$$\mathsf{A}_\gamma(P_q) = P_q \cup \{i \in C_q : s_q(i) \leq \gamma\}. \quad (8)$$

Thus a requested deletion remains deleted only when its item score strictly exceeds γ. This single rule represents all 2^{K_q} proposals implicitly. Setting $P_q = \varnothing$ retains exactly the items whose deletion is not certified; it is therefore the smallest authorized retained context, or equivalently the deletion-maximal authorized action.

The online scan uses item scores and a frozen threshold, without audit labels or subset search. Ties remain retained. If no calibration rank qualifies, $\gamma = +\infty$ restores the registered complete frame. Hard-budget abstention preserves all-question harm containment; risk conditional on answering requires the selective audit below.

Calibration, once per registered scorer and question distribution

1. Fit the item deletion scorer on proper-training questions. Fix the candidate factory, support loss, risk target, and calibration confidence.
2. For each calibration question, score the items and retain only the maximum over its annotated support, $Z(q) = \max_{i \in S_q} s_q(i)$.
3. Choose the least admissible rank in Eq. 6, or the finite-frame rank in Eq. 18. Freeze its threshold γ and the execution rule.

Execution, for each new question without support labels

1. Score the registered item frame C_q and obtain any proposed retained set P_q.
2. Scan the frame once, restoring each item with $s_q(i) \le \gamma$. Return $P_q \cup \{i : s_q(i) \le \gamma\}$ in the registered evidence order.
3. Execute the registered answer or fallback rule on this complete context. A hard budget may reject the context; it may not silently truncate it.

Figure 1: **From support annotations to a label-blind authorizer.** Support labels enter calibration, not execution. The $O(K_q)$ authorization scan is separate from the cost of computing the learned item scores.

Random-trace selective audit. Let $(T_i, A_i, L_i)_{i=1}^n$ be independent audit traces, where $A_i = \mathbb{I}[\pi(T_i) \neq \perp]$ and $L_i = \ell(Q_i, T_i, \pi(T_i))$; registration requires $L_i = 0$ when $A_i = 0$. Write $M = \sum_i A_i$ and $X = \sum_i L_i$. For $0 \le X < M$, let $U_{\mathrm{CP}}(X, M; \delta)$ be the one-sided Clopper–Pearson upper endpoint, the unique u satisfying

$$\sum_{j=0}^{X} \binom{M}{j} u^j (1-u)^{M-j} = \delta, \tag{9}$$

and set $U_{\mathrm{CP}} = 1$ when $X = M$, including $M = 0$. This is the one-sided Clopper–Pearson bound (Clopper & Pearson, 1934). A policy is admitted only when $U_{\mathrm{CP}} \le \alpha$ and any registered coverage floor is met. If one of $|\Pi|$ policies fixed before the audit is selected using these same labels, replace δ by $\delta/|\Pi|$ for each policy; a policy chosen on separate development data incurs no such audit-time multiplicity.

For an enumerated public frame, calibration is a simple random sample without replacement from the post-training frame. The exact hypergeometric rank in Appendix A.3 controls the entire post-calibration complement, rather than treating its members as fresh i.i.d. draws.

Both routes execute without audit labels. Their guarantees apply to the calibrated policy–loss family and deployment distribution, including selection within a simultaneously certified menu.

5 Theoretical Guarantees

The theorem connects the two method routes. Its first three parts characterize the exact envelope for a family of deletion actions and the cost of computing it. The fourth separates support missing from the retrieved candidates from support removed by later deletion, then bounds the coverage of any policy restricted to those candidates. Corollaries 1–2 provide the corresponding finite-sample calibration statements. The max–min representation uses classical lattice polynomials (Dukhovny & Marichal, 2008); the oracle bounds use decision-tree arguments (Buhrman & de Wolf, 2002).

Theorem 1 (exact envelope, computation, and retrieval ceiling). *(i) Power-set envelope.* Let A_γ be Eq. 8. For every registered item frame and score vector,

$$\max_{P_q \subseteq C_q} \mathbb{I}[S_q \not\subseteq \mathsf{A}_\gamma(P_q)] = \mathbb{I}[S_q \not\subseteq \mathsf{A}_\gamma(\varnothing)] = \mathbb{I}[Z(q) > \gamma]. \tag{10}$$

This covers all 2^{K_q} proposals, even when selected after observing scores. The empty proposal is deletion-maximal among label-blind actions preserving this inclusion for every support set and is

constructed in $O(K_q)$ time; $Z(q)$ is its pointwise-smallest scalar envelope by the minimality argument in Section 3.

(ii) Structural characterization. Let $h_q(D)$ be any binary harm of the deleted set, with $h_q(\varnothing) = 0$, and let $\mathcal{W}_q$ contain its inclusion-minimal harmful sets. For $H_\gamma = \{i : s_q(i) > \gamma\}$,

$$\max_{P_q \subseteq C_q} h_q(H_\gamma \setminus P_q) = \mathbb{I}\left[\max_{W \in \mathcal{W}_q} \min_{i \in W} s_q(i) > \gamma\right], \tag{11}$$

with an empty maximum equal to $-\infty$. This equals $\mathbb{I}[\max_{i \in S} s_q(i) > \gamma]$ for some score-independent S, for every score vector and threshold, iff every minimal witness is a singleton. Given the witnesses, their envelope costs $O(\sum_{W \in \mathcal{W}_q} |W|)$ to evaluate. Support omission has exactly the witnesses $\{\{i\} : i \in S_q\}$, yielding part (i) without enumerating proposals.

(iii) Tight oracle separation. Suppose the scores are known and h_q is accessible through exact queries $D \mapsto h_q(D)$. If h_q is monotone and the scores have m distinct values, its exact envelope can be found in $\lceil \log_2(m+1) \rceil$ queries after sorting. For distinct scores this worst-case query bound is optimal. Without monotonicity or other structural information, exact deterministic envelope computation requires $2^{K_q} - 1$ queries in the worst case. Even randomized algorithms correct with probability at least $1 - \varepsilon$ for every loss need at least $(1 - 2\varepsilon)(2^{K_q} - 1)$ worst-case queries, for $0 \le \varepsilon < 1/2$. These bounds concern the exact smallest envelope, not a loose conservative bound.

(iv) Incomplete retrieved frames. Let $C(T)$ be a finite query-dependent candidate set, $B = \mathbb{I}[S \not\subseteq C(T)]$, and $Z_C = \max_{i \in S \cap C(T)} s(i)$. For finite γ, apply Eq. 8 inside $C(T)$. Then

$$\max_{P \subseteq C(T)} \mathbb{I}[S \not\subseteq \mathsf{A}_\gamma(P)] = B + (1 - B)\mathbb{I}[Z_C > \gamma]. \tag{12}$$

More generally, every possibly randomized answer-or-abstain rule whose returned context lies inside $C(T)$ has coverage c and selective support risk r_s satisfying

$$c(1 - r_s) \le 1 - \Pr(B = 1), \qquad r_s \le \alpha_s < 1 \implies c \le \min\left\{1, \frac{1 - \Pr(B = 1)}{1 - \alpha_s}\right\}. \tag{13}$$

This ceiling is sharp for an oracle that observes retrieval completeness and covers arbitrary downstream gate tuning within the same candidate set.

Proof idea. Maximizing over proposals takes the upward closure of harm, whose minimal witnesses give part (ii). Monotonicity reduces oracle access to one prefix transition; otherwise an unqueried subset can hide the only harmful action. For part (iv), unretrieved support is absent from every output; on complete frames part (i) applies. A support-complete returned answer requires $B = 0$, giving Eq. 13. Appendices A.2 and D.7 give the proofs and exhaustive checks.

For $K_q = 20$, an unrestricted exact harm oracle can require 1,048,575 calls; a monotone oracle needs at most five. Required-support annotations remove those oracle calls entirely: compute $Z(q)$ and authorize by one item scan. Answer error need not be monotone in deleted evidence, so evaluating only the largest authorized deletion can miss a harmful smaller deletion. Corollary 2 instead audits the frozen executed policy using one loss observation per trace. Both audit complete actions, but the subset-family guarantee covers all proposals, whereas the open-trace audit covers a fixed deployed policy.

Corollary 1 (calibration of the power-set envelope). If γ is chosen by Eq. 6 from conditionally i.i.d. question maxima, then with calibration confidence $1 - \delta$, the fresh-question harm probability is at most α. If calibration is instead $C \sim \mathrm{SRSWOR}_n(F)$ and k is chosen by Eq. 18, then with probability at least $1 - \delta$,

$$\frac{1}{R} \sum_{q \in A} \mathbb{I}[Z_q > \gamma] \le \alpha. \tag{14}$$

Here $A = F \setminus C$ and $R = |A|$. Both ranks are least conservative within their order-statistic classes. Any executed action whose harm is pointwise contained in the envelope inherits its guarantee; Appendix A treats nested prefixes, ties, and imperfect fallback.

Corollary 2 (simultaneous exact open-trace certification). Let $(Q_i, T_i)_{i=1}^n$ and a fresh (Q, T) be i.i.d., where $T = G(Q, \mathcal{W}, U)$ includes the complete adaptive retrieval trace and all generator randomness. Trace lengths and action identities may be arbitrary and question-dependent. Fix before the audit a measurable policy $\pi(T) \in T \cup \{\perp\}$ and binary audit loss ℓ satisfying $\ell(Q, T, \perp) = 0$. Let M and X be the answered and harmful-answer counts on the audit sample. Then, with probability at least $1 - \delta$ over that sample,

$$R_{\mathrm{end}}(\pi) \le R_{\mathrm{sel}}(\pi) \le U_{\mathrm{CP}}(X, M; \delta). \tag{15}$$

The statement requires independence only across complete traces, not among their internal steps, candidates, or retrievers. More generally, fix a finite library Π and J losses before audit and set

$$U_{\pi j} = U_{\mathrm{CP}}\left(X_{\pi j}, M_\pi; \frac{\delta}{J|\Pi|}\right). \tag{16}$$

With probability at least $1 - \delta$, all $J|\Pi|$ inequalities hold simultaneously, allowing selection of the certified policy with largest observed M_π/n. Independent development selection removes the $|\Pi|$ factor. For incomplete binary outcomes, replacing $X_{\pi j}$ by observed errors plus all missing answered outcomes preserves validity: the replacement dominates the true count pathwise and U_{CP} is nondecreasing. No missing-at-random assumption is required. Appendix A proves these standard exact-binomial consequences and the endpoint's optimality within deterministic nondecreasing bounds.

6 Experiments

The experiments mirror the paper's two settings and then connect them. The fixed-frame audits test whether complete-action calibration controls support omission; the matched MuSiQue experiment separates calibration from scoring; the open-retrieval audit tests whether changing the candidate factory can exceed the ceiling for BM25-based gates. Results are reported as separate audits because their sampling units and confidence events differ.

6.1 Protocol

HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) provide annotated multi-hop support. Development and test roles are separated by public question ID. Labels are opened only for their registered role. Fixed-frame tests compare authorization rules on complete candidate pools with sufficient fallback. Open-retrieval tests compare complete systems whose candidates are constructed per query from the October-2017 Wikipedia projection or the fixed native 2Wiki corpus. Appendices B–D give the protocols and additional audit details for the claims reported here.

The fixed-frame evaluation uses 5,000 disjoint MuSiQue calibration/test questions and complete 7,405-question HotpotQA and 12,576-question 2Wiki frames. Open-trace tests comprise three 10,000-question HotpotQA roles, a 20,000-question strict replication, and a fresh 10,000-question post-answer audit, all ID-disjoint. The six-policy 2Wiki audit fixes six policies before 20,000 independent draws from unused question IDs. MuSiQue fixes candidates and Qwen2.5-7B decoding across full evidence, our authorizer, native Conformal-Embedding, and exact-cardinality Qwen3 dense retrieval. Frozen-context checks also use Mistral-7B-Instruct-v0.3 (Jiang et al., 2023); open-corpus answer audits include Qwen2.5-14B. The Qwen2.5 and Qwen3-Embedding reports document the model families used here (Qwen Team, 2024; Zhang et al., 2025).

The MuSiQue scorer predicts paragraph support from Qwen3 dense similarity, BM25, Qwen2.5-7B necessity scores, lexical overlap, and inter-passage connectivity. Connectivity features summarize title mentions, token overlap, and the retrieval ranks of linked passages; they are computed from candidate text rather than support labels. Model selection uses 5,000 previously opened development questions, divided into 3,000 fitting, 1,000 calibration, and 1,000 evaluation questions. A histogram gradient-boosted classifier is selected and refitted on all 5,000 development questions before final roles are allocated. It uses at most 31 leaves and at least 20 samples per leaf, with a 600-iteration limit and learning rate .03. Its deletion score is one minus the predicted support probability. The final calibration and test IDs exclude this entire development pool. Thus the learned arm differs from

native embedding through its score construction, while the factorial replay holds each resulting score vector fixed when changing the calibration rule.

Conformal-Embedding uses Qwen3-Embedding-8B revision `b38b5852`, its official retrieval instruction, last-nonpadding-token pooling, L2 normalization, and the positive-calibration quantile. All four MuSiQue arms receive the same 39,992 candidate paragraphs. The dense arm ranks them by Qwen3 similarity and matches our retained count separately for every question, rather than matching only the mean budget. The frozen generation comparison uses 50,000 question-paired bootstrap draws with seed 2026090204. Its ledger contains 7,359 unique generation states, each with two byte-identical token replicas; replay reconstructs all 8,000 method–question rows.

6.2 Calibrating complete actions preserves support

Table 1 isolates the calibrated object on 500 HotpotQA and 500 2Wiki questions. Each of five question-grouped folds uses 200/200/100 training/calibration/test questions per dataset, with the scorer and candidates fixed across rules. These are pooled out-of-fold measurements. Candidate pooling exceeds the target in every fold. Retriever-wise Bonferroni is family-valid but nearly vacuous; the family-valid complete-action maximum retains 18.10% safe-early coverage at $\alpha = .05$ and 40.37% at $\alpha = .10$.

Table 1: **The audit unit determines validity and utility.** The scorer, candidates, and five question-grouped folds are fixed. Q-harm is measured over 1,000 question clusters; safe early coverage is measured over 3,000 retriever–question decisions. Family-valid refers to the calibration rule's target, while the reported rates are out-of-fold measurements.

α	Calibration rule	Family-valid?	Q-harm	Safe early	Over budget
5%	Candidate pooled	no	13.80%	61.70%	5/5
	Retriever Bonferroni	yes	0.00%	0.00%	0/5
	Complete-action max	**yes**	**1.00%**	**18.10%**	**0/5**
10%	Candidate pooled	no	24.80%	77.43%	5/5
	Retriever Bonferroni	yes	0.50%	1.73%	0/5
	Complete-action max	**yes**	**5.00%**	**40.37%**	**0/5**

The prospective complete frames give the corresponding frozen test. HotpotQA has 179 harmful questions in its 5,405-question complement (cap 270), and 2Wiki has 398/10,576 (cap 528), with zero containment violations. Appendix B and Table 7 give the full audit.

6.3 Uniform authorization enables scorer-driven compression

The final MuSiQue protocol applies Eq. 8 with $P_q = \varnothing$ rather than fitting a separate empirical controller. On 2,000 held-out questions, its 83 support omissions give a one-sided 95% Clopper–Pearson upper bound of 4.96%, below the frozen 5% target. It retains 10.5185 of 19.996 candidate paragraphs on average: 47.40% point reduction with a 46.70% question-bootstrap lower bound, far above the frozen 25% gate.

Table 2: **Theorem-aligned MuSiQue confirmation** on the same 2,000 questions, candidate paragraphs, and Qwen2.5-7B decoder. Qwen3 dense is forced to match $\mathsf{A}_\gamma(\varnothing)$'s retained count on every question (zero budget mismatches). F1 and EM are absolute percentages.

Method	Omit ↓	Red. ↑	Kept	F1 ↑	EM ↑
Full evidence	0.00	0.00	20.00	42.67	33.30
$\mathsf{A}_\gamma(\varnothing)$ (ours)	**4.15**	**47.40**	10.52	**47.94**	**37.75**
Native Conformal-Embedding	10.05	13.75	17.25	42.59	32.85
Qwen3 dense, same $\lvert\mathsf{A}\rvert$	42.35	47.40	10.52	41.00	31.45

Native Conformal-Embedding (Chakraborty et al., 2026) uses its official Qwen3 embedding rule with the same candidates and answer decoder. Its complete-context omission is 10.05%; the paired

native-minus-ours omission lower endpoint is $+4.35$ points. The exact-budget Qwen3 dense arm omits support in 42.35%. Against full evidence, our F1/EM changes are $+5.27/+4.45$ points, with one-sided 97.5% lower bounds $+3.65/+2.80$. All five frozen gates pass. Appendix C records the dual-token replay and 50,000 paired bootstrap draws.

Table 3 separates the effects of scoring and calibration on the same 3,000/2,000 calibration/test split. Each frozen scorer uses both native positive-pair and complete-action maximum calibration. With either scorer fixed, complete-action calibration lowers question omission from about 10% to about 4% by retaining more evidence. For the learned scorer, EM falls by 2.15 points (paired 95% interval $[-3.55, -.75]$). With complete-action calibration fixed, learned scores instead improve EM by 4.70 points ($[3.10, 6.30]$), reduce mean evidence from 18.68 to 10.52 paragraphs, and reduce prompt tokens by 37.71%. The support guarantee comes from calibration; the compression and answer-quality gains depend on the scorer.

Table 3: **Scorer $\times$ calibration object with matched answers.** Same 2,000 questions, candidates, and decoder; thresholds use only 3,000 calibration questions. Nominal $\alpha = .05$; maximum calibration also uses $\delta = .05$. Omission and F1/EM are percentages; prompt cost is millions of tokens. All arms answer every question.

Frozen scorer	Calibration	Omission	Kept	F1 / EM	Prompt M
Qwen3 embedding	Positive pair	10.05	17.25	42.59 / 32.85	4.487
Qwen3 embedding	Action maximum	3.70	18.68	42.51 / 33.05	4.839
Learned support	Positive pair	10.20	7.61	49.94 / 39.90	2.266
Learned support	Action maximum	4.15	10.52	47.94 / 37.75	3.014

Both calibrators span the same global-threshold class for a fixed scorer. A separate exhaustive support-only sweep gives embedding unrestricted hindsight tuning over every test-score breakpoint; it still has no positive feasible coverage under cap 15 at empirical omission at most 5%. Appendix C.1 separates this 1,597,740-point support oracle from the fully generated calibration grid.

Both scorers receive the same candidates, decoder, calibration roles, 24 risk levels, and 20 hard budgets. A context exceeding its cap triggers abstention. The answer grid covers all 1,920 registered points. Table 8 reports the best correct-answer yield at each cap with simultaneous support upper bound at most 5%. At cap 15, the learned scorer yields 33.10 correct answers per 100 questions at 82.10% coverage; the embedding grid has no feasible point. At cap 20, yields are 37.80 and 33.55, with 3.21M versus 5.09M prompt tokens. The table reports both answer-error and support bounds for these descriptive grid optima.

We also optimize mixtures of the registered thresholds, budget rules and abstention under equal prompt-cost and empirical risk constraints. The embedding comparator receives both calibration menus. Independent primal–dual checks verify 576 programs reconstructed from 192,000 answers (Appendix C.2). At 25% answer error and 5% support omission, correct-answer yield is at most 0.3472% for learned scores and zero for embedding; the larger efficiency gap occurs at a 75% answer-error limit. This motivates testing whether changing retrieval can exceed the fixed-candidate coverage ceiling.

6.4 Changing retrieval exceeds the subset-gate ceiling

Post-answer gating can improve selective answer quality without changing retrieval. On HotpotQA, the strict 20,000-question audit does not meet the 25% EM-error target: its upper bound is 25.91%. A policy fixed before a fresh 10,000-question audit meets both risk targets at 42.54% coverage, with support-omission and EM-error upper bounds of 4.20% and 22.01%. Within this fresh cohort, a paired replay of the base and post-answer gates shows that the latter rejects 195 incorrect and two correct answers. Appendices D.1–D.3 report the protocols, estimates, and costs. EM uses normalized strings and registered aliases; risk is conditional on returned answers, whereas coverage divides by all allocated questions.

To test the retrieval ceiling on 2Wiki, we compare BM25 and title-linked retrieval with no gate, occurrence gating, and a frozen joint gate. All six policies share the question draws, Top-20 budget, Qwen2.5-14B prompt, and greedy decoder. The linked selector is fitted on 19,432 opened

development IDs and frozen before 20,000 i.i.d. draws with replacement from unused IDs. Its 31 features, 70 interactions, ridge strength 100, and answer/support cutoffs $.40/.075$ are fixed in advance. A $.025/18$ allocation covers both conditional risks and coverage for all six policies (Table 4). Only linked joint gating meets all three targets: 8,819 answered draws contain 808 EM errors and 294 support omissions. Its coverage exceeds BM25 joint gating by 33.29 points (descriptive paired 95% interval $[32.63, 33.95]$). Warm-service replay measures 1.446 seconds per query for linked joint gating versus 1.217 for BM25, an 18.8% latency increase including retrieval, generation and selection but excluding offline training and initialization (Appendix D.9). Appendix D.5 reports the matched development comparison.

Table 4: **Fresh 2Wiki menu and certified mixture.** Six frozen policies share 20,000 draws and one simultaneous 97.5% event; the post-audit mixture reuses this event without new generation. UB/LB: upper/lower bounds. Targets: 25% answer error, 5% support omission, 40% coverage.

Retriever	Gate	Observed cover	Cover LB	Answer UB	Support UB
BM25	None	100.00%	99.97%	50.31%	66.41%
BM25	Occurrence	77.66%	76.77%	37.57%	57.38%
BM25	Joint	10.81%	10.16%	7.29%	3.28%
Linked	None	100.00%	99.97%	26.92%	11.44%
Linked	Occurrence	95.26%	94.79%	23.34%	10.06%
Linked	Joint	44.10%	**43.04%**	**10.12%**	**3.95%**
Linked	92:8 mixture	–	**47.18%**	**12.34%**	**4.973%**

BM25 misses complete support on 13,082/20,000 draws. Its one-sided 97.5% lower endpoint is $b_L = .64746$. Intersecting this event with the six-policy 97.5% event gives a joint 95% certificate. Theorem 1(iv) caps every BM25 subset gate with support risk at most 5% at $U_C = (1 - b_L)/.95 = 37.11\%$ coverage, hence at most 37.11% correct-answer yield. This includes randomized, oracle and arbitrarily tuned gates, not just the registered menu. The frozen linked gate instead has coverage at least 43.04% and correct yield at least $L_c(1 - U_a) = 38.69\%$: simultaneous advantages of at least 5.93 and 1.58 points, while meeting both strict risk targets. Appendix D.8 gives the uniform comparison.

Table 5: **Strict-risk separation from every BM25 subset gate.** Joint confidence is at least 95%. Row 1 bounds all gates on the original BM25 Top-20 candidates. Both linked policies meet 25% answer error and 5% support omission. Yield is correct answers per allocated query.

Policy or class	Coverage	Correct yield	Answer UB	Support UB
All BM25 subset gates	$\leq 37.11\%$	$\leq 37.11\%$	–	5% required
Frozen linked joint	$\geq 43.04\%$	$\geq 38.69\%$	10.12%	3.95%
Derived linked mixture	$\geq 47.18\%$	$\geq 41.41\%$	12.34%	4.973%

The frozen linked policy exceeds the original candidate factory's coverage ceiling while meeting both risk targets. The comparison isolates the risk–coverage consequence of changing retrieval.

7 Conclusion

The witness envelope makes complete-action authorization tractable. Matched experiments separate calibrated support preservation from scorer-driven compression. Under both risk targets, linked retrieval's 43.04% coverage lower bound exceeds the 37.11% ceiling for every admissible gate on the original BM25 candidates; a certified mixture raises the floor to 47.18%. Together, these results give a structural account of risk-controlled evidence selection: harmful-action structure determines tractable authorization, while retrieval completeness limits achievable coverage. The resulting bounds provide a practical diagnostic for identifying coverage targets that require changing retrieval itself.

Reproducibility Statement

Appendix A states the assumptions and gives the full proofs; Appendices B–D specify the audit protocols and report results supporting the main claims. Independent programs reconstruct the evaluated actions, replay saved generation tokens, and recompute metrics and frozen resampling streams. The supplementary archive includes these programs, per-question records, frozen protocols, and file checksums.

AI Use Statement

Generative AI tools assisted in retrieving and discovering related literature and in polishing the manuscript for clarity and presentation. The authors independently reviewed the retrieved references and the final manuscript and take responsibility for all submitted content, including AI-assisted text.

References

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

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

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning to retrieve reasoning paths over Wikipedia graph for question answering. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgVHkrYDH.

Harry Buhrman and Ronald de Wolf. Complexity measures and decision tree complexity: A survey. *Theoretical Computer Science*, 288(1):21–43, 2002. doi: 10.1016/S0304-3975(01)00144-X. URL https://homepages.cwi.nl/~rdewolf/publ/qc/dectree.pdf.

Debashish Chakraborty, Eugene Yang, Daniel Khashabi, Dawn Lawrie, and Kevin Duh. Principled context engineering for RAG: Statistical guarantees via conformal prediction. In *Advances in Information Retrieval*, volume 16484 of *Lecture Notes in Computer Science*, pp. 537–546. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-21300-6_45. URL https://arxiv.org/abs/2511.17908.

Lu Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. Controlling risk of retrieval-augmented generation: A counterfactual prompting framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2380–2393. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.133. URL https://aclanthology.org/2024.findings-emnlp.133/.

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

Srikanth Devarakonda, Rajesh Lingam, and Vagdevi Challa. Confidence-gated RAG for adaptive retrieval in sequential agents. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. URL https://openreview.net/forum?id=5BOaSaHY6t.

Alexander Dukhovny and Jean-Luc Marichal. System reliability and weighted lattice polynomials. *Probability in the Engineering and Informational Sciences*, 22(3):373–388, 2008. doi: 10.1017/S0269964808000223. URL https://doi.org/10.1017/S0269964808000223.

Vojtech Franc, Daniel Prusa, and Vojtech Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL https://www.jmlr.org/papers/v24/21-0048.html.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. When to retrieve during reasoning: Adaptive retrieval for large reasoning models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026. doi: 10.1145/3805712.3809722.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Ansh Kamthan. AB-RAG: Adaptive budgeted retrieval-augmented generation for reliable question answering. *arXiv preprint arXiv:2606.29090*, 2026.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Adrian Kieback, Uyiosa Philip Amadasun, Aman Chadha, and Aaron Elkins. TASR: Training-free adaptive stopping for iterative retrieval. *arXiv preprint arXiv:2606.13814*, 2026. URL https://arxiv.org/abs/2606.13814. Accepted at the Agent4IR Workshop, KDD 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25846–25862. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1185. URL https://aclanthology.org/2026.acl-long.1185/.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3799–3821. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Jeffery Opoku and David Banahene. ToolChain-CRC: Conformal risk control for agentic AI under retrieval and tool-use drift. *arXiv preprint arXiv:2606.18467*, 2026. URL https://arxiv.org/abs/2606.18467.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-RAG: Value-based retrieval control for iterative RAG. *arXiv preprint arXiv:2510.14337*, 2025. URL https://arxiv.org/abs/2510.14337. NeurIPS 2025 MTI-LLM Workshop.

Pin Qian, Su Wang, Chong Peng, Junxian You, Lifei Liu, Haoran Yu, Yihang Chen, and Xiaochong Jiang. When should active RAG retrieve? a budget-aware evaluation of utility, calibration, and cost. *arXiv preprint arXiv:2607.24010*, 2026. URL https://arxiv.org/abs/2607.24010. Accepted at the KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026. Submitted to IEEE PRAI 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. doi: 10.48550/arXiv.2212.03533. URL https://arxiv.org/abs/2212.03533.

Ting Wang, Yuanjie Shi, Yan Yan, and Huan Zhang. Inference-time conformal reasoning with valid factuality control for large language models. *arXiv preprint arXiv:2606.08831*, 2026a. URL https://arxiv.org/abs/2606.08831. Accepted at ICML 2026.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: Risk control for reasoning on a compute budget. *arXiv preprint arXiv:2602.03814*, 2026b. URL https://arxiv.org/abs/2602.03814. Accepted at ICML 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023. doi: 10.48550/arXiv.2309.07597. URL https://arxiv.org/abs/2309.07597.

Roy Xie, Junlin Wang, Paul Rosu, Chunyuan Deng, Bolun Sun, Zihao Lin, and Bhuwan Dhingra. Knowing when to stop: Efficient context processing via latent sufficiency signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1376. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3afa7f9b7ac8da474b3d915570d58291-Abstract-Conference.html.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EMHoBG0avc1.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Zihan Zhu, Shayan Kiyani, George Pappas, and Hamed Hassani. Conformal risk-averse decision making with action conditional guarantee. *arXiv preprint arXiv:2606.05551*, 2026. URL https://arxiv.org/abs/2606.05551.

A Proof Details

This appendix contains proofs, audit protocols, and additional results supporting the main claims. Section A.2 proves the structural and oracle results; Section D.7 proves the retrieval ceiling. Sections C and D.6 report the corresponding experiments.

A.1 Proof that marginal certificates do not compose

Let $U \sim \mathrm{Uniform}[0,1)$ and define E_r as a circular interval of length α, with the m interval starts spaced to cover as much of the unit circle as possible. Each event has probability α. If $m\alpha \le 1$, choose disjoint intervals and obtain $\Pr(\cup_r E_r) = m\alpha$. If $m\alpha > 1$, choose the intervals so that they cover the circle and obtain union probability one. Hence the largest union compatible with the marginal constraints is $\min\{1, m\alpha\}$, which exceeds α whenever $m > 1$ and $\alpha > 0$. Bonferroni allocation at α/m controls the union but may reduce coverage. Equation 5 instead calibrates the union event directly by taking the maximum after all co-occurring actions and sources have been assembled for one question.

A.2 Proof of Theorem 1

The power-set clause is proved in the paragraph "Exact power-set reduction and deletion maximality" below, together with its calibration consequences. For part (ii), define the upward closure $\bar{h}(H) = \max_{D \subseteq H} h(D)$. As P ranges over the power set, $H_\gamma \setminus P$ ranges over every subset of H_γ. Hence the proposal maximum is $\bar{h}(H_\gamma)$, whether or not h is monotone. By finiteness, $\bar{h}(H) = 1$ iff H contains an inclusion-minimal harmful set: a harmful subset contains a minimal one, and a contained minimal witness is itself an available harmful deletion. A witness W is contained in H_γ exactly when $\min_{i \in W} s(i) > \gamma$, proving Eq. 11. Scanning supplied witness incidences evaluates its max–min value. The evaluation cost is measured in the supplied witness incidences.

For necessity of singleton witnesses, choose scores one on an arbitrary D and zero elsewhere, with $\gamma = 1/2$. An item-maximum representation over a fixed S would imply $\bar{h}(D) = \mathbb{I}[D \cap S \neq \varnothing]$ for every D. The minimal true sets of $\bar{h}$ are precisely the minimal harmful sets of h, so they must be $\{\{i\} : i \in S\}$. Conversely, such witnesses reduce the max–min expression to $\max_{i \in S} s(i)$. The zero-harm function uses $S = \varnothing$. As an interacting example, if either of two passages alone suffices, the sole minimal harmful deletion is $\{1, 2\}$ and its exact envelope is $\min(s(1), s(2))$, not $\max(s(1), s(2))$. This identifies which loss representation must be supplied to calibration.

The exhaustive four-item regression enumerates all 167 monotone binary harms satisfying $h(\varnothing) = 0$. Exactly 16 have singleton-only minimal witnesses. Across every score ordering, every threshold cut, and all 16 proposals, the max–min identity passes 320,640 checks with zero mismatches. This is a finite mathematical regression, separate from the empirical retrieval experiments.

Part (iii): optimal monotone oracle search. Group equal scores and write the distinct levels as $t_1 > \cdots > t_m$. Let $D_j = \{i : s(i) \ge t_j\}$, with $D_0 = \varnothing$. Monotonicity makes $h(D_0), \ldots, h(D_m)$ a binary sequence beginning at zero with at most one transition. Define $j^* = \min\{j : h(D_j) = 1\}$, using $m+1$ if none is harmful. For a monotone loss, $\bar{h} = h$, so the exact envelope is t_{j^*} if $j^* \le m$ and $-\infty$ otherwise. Binary search over the $m+1$ transition positions needs at most $\lceil \log_2(m+1) \rceil$ oracle calls. Grouping ties is necessary because no strict score threshold can separate tied items. Sorting costs $O(K \log K)$ comparisons; the theorem counts the subsequent loss-oracle calls separately.

For optimality with distinct scores, consider the $K+1$ monotone losses $h_r(D) = \mathbb{I}[|D| \geq r]$, $r = 1, \ldots, K+1$. They have different envelopes, including $-\infty$ for $r = K+1$. A correct binary-query decision tree must distinguish all of them. Depth d has at most 2^d leaves, so $d \geq \lceil \log_2(K+1) \rceil$, matching the algorithm even when queries are not restricted to score prefixes.

Fix any finite score vector and suppose $h(\varnothing) = 0$ is known. On the all-safe oracle $h_0(D) = 0$, a deterministic algorithm that omits a nonempty deletion D^* has exactly the same transcript on $h_{D^*}(D) = \mathbb{I}[D = D^*]$. The first envelope is $-\infty$ and the second is the finite value $\min_{i \in D^*} s(i)$. Thus exact computation requires every one of the $2^K - 1$ nonempty queries on this input. Querying them all and taking the largest harmful bottleneck attains the bound.

For a randomized algorithm, couple the executions on h_0 and h_{D^*} using the same internal coins. Until D^* is queried, their transcripts and outputs coincide. Let Q_{D^*} be the event that D^* is queried on the all-safe execution, and let F_0, F_{D^*} be their respective error events. Since the correct envelopes differ, $Q_{D^*}^c \subseteq F_0 \cup F_{D^*}$. Uniform error at most ε implies $\Pr(Q_{D^*}) \geq 1 - 2\varepsilon$. Summing over all nonempty D^* gives expected queries on h_0 at least $(1 - 2\varepsilon)(2^K - 1)$, and therefore the same lower bound on worst-case queries. This applies the classical hidden-OR adversary (Buhrman & de Wolf, 2002) to computing the exact envelope.

Take scores $(3, 2, 1)$ and let only deletion $\{1, 2\}$ be harmful. At $\gamma = 0$, $H_\gamma = \{1, 2, 3\}$ is harmless, but the proposal retaining item 3 executes the harmful deletion. Its envelope is 2. Thus nonmonotone answer loss cannot generally be represented by the loss of H_γ alone. If the full context is already harmful, $h(\varnothing) = 1$, the empty deletion is available at every threshold and the proposal-uniform envelope is $+\infty$. The oracle separation uses the safe-full-context premise. The open-trace audit measures the executed policy directly, including full-context errors.

The expanded regression enumerates all $2^{15} = 32,768$ binary losses on four items with safe empty deletion, all 24 strict score orders, and all five threshold cuts. Direct maximization over proposals agrees with the minimal-witness envelope in 3,932,160 event checks. A separate implementation uses the Boolean subset-zeta transform to reconstruct $\bar{h}$, obtaining the same count. The original monotone cases form a 167-loss subset. Binary search agrees with full enumeration on 4,008 monotone/order pairs and 2,672 tied-score cases. In 1,075,176 loss/order/cut combinations, evaluating only the largest deletion misses an available harmful subset. These are exhaustive finite checks of the stated mathematics.

A.3 Proof of Corollary 1

Conditional-i.i.d. calibration clause. Fix one dataset group chosen before calibration and condition on proper training, which fixes the scorer and registered candidate factory without using calibration harm labels. By assumption, the n calibration maxima and a fresh test maximum are i.i.d. draws from this group's registered distribution. Let F be their common CDF, including any mass at $-\infty$, and define $q_{1-\alpha} = \inf\{x : F(x) \geq 1 - \alpha\}$. If the conditional upper-tail probability above $M_{(k)}$ exceeds α, then $M_{(k)} < q_{1-\alpha}$. Moreover, $\Pr\{M < q_{1-\alpha}\} = F(q_{1-\alpha}-) \leq 1 - \alpha$. Therefore

$$\Pr_{\text{cal}}\big\{\Pr(M > M_{(k)} \mid \text{cal}) > \alpha\big\} \leq \Pr\{\text{Binomial}(n, 1-\alpha) \geq k\} \leq \delta_g.$$

This argument permits discrete scores and ties; strict deployment thresholding is essential because it uses the upper tail $\Pr(M > M_{(k)})$. It directly controls $B_\gamma(q) = \mathbb{I}[M(q) > \gamma]$, which contains the harmful-selection event of any direct threshold selector. The maximum is computed after collecting all actions and retrievers that can co-occur for one question, so their number and dependence never enter the bound. For mutually exclusive dataset stratum g, let γ_g be its calibrated threshold. If all strata share one registered real-valued scorer and numerical scale, deploy $\gamma = \max_h \gamma_h$; otherwise use γ_g only in stratum g. In either case $\Pr_g(M > \gamma_g) \leq \alpha$ whenever the stratum-g certificate holds, and the common-threshold version is no less conservative. A union bound over calibration-failure events gives simultaneous confidence at least $1 - \sum_g \delta_g$. Splitting δ does not control a population-risk union of co-occurring action groups; those actions must be included in the same question maximum or assigned separate α budgets.

For optimality within calibration order statistics, take a continuous null distribution, such as a uniform transform. Then the preceding inequality is exact:

$$\Pr_{\text{cal}}\big\{\Pr(M > M_{(k)} \mid \text{cal}) > \alpha\big\} = \Pr\{\text{Binomial}(n, 1-\alpha) \geq k\}.$$

Every smaller index that fails Eq. 6 therefore fails the requested confidence for this distribution. The smallest admissible k is consequently the lowest, and hence least conservative, uniformly valid order-statistic threshold. If no index exists, $\gamma = +\infty$ accepts no risk action and is valid trivially.

Exact power-set reduction and deletion maximality. Fix a question, a proposal $P_q \subseteq C_q$, and the calibrated threshold γ. If the authorized output is harmful, there is an item $i \in S_q$ with $i \notin \mathsf{A}_\gamma(P_q)$. Equation 8 then implies both $i \notin P_q$ and $s_q(i) > \gamma$. Hence

$$Z(q) = \max_{j \in S_q} s_q(j) \geq s_q(i) > \gamma,$$

so every proposal's harmful event is contained in $\{Z(q) > \gamma\}$. Conversely, if $Z(q) > \gamma$, choose $i^\star \in S_q$ with $s_q(i^\star) = Z(q)$. The empty proposal retains exactly the items with score at most γ, so $i^\star \notin \mathsf{A}_\gamma(\varnothing)$ and its authorized output is harmful. Since $\mathsf{A}_\gamma(\varnothing) \subseteq \mathsf{A}_\gamma(P_q)$ for every proposal, the empty proposal attains the family maximum. This proves all three equalities in Eq. 10. The argument holds simultaneously for every one of the 2^{K_q} proposals and for any adaptive rule that chooses among them. The order-statistic calibration is applied only to $Z(q)$, so neither proposal cardinality nor within-question dependence enters the probability bound.

For maximality, take any item i with $s_q(i) \leq \gamma$. Suppose a label-blind rule could delete i while preserving Eq. 10 for every possible support set. Choose the compatible support set $S_q = \{i\}$. The rule's output is harmful, but $Z(q) = s_q(i) \leq \gamma$, contradicting the required inclusion. Thus every uniformly valid itemwise rule must retain all items with score at most γ. Equation 8 with $P_q = \varnothing$ retains exactly those items, so no other uniformly admissible output can delete a strict superset. One scan over the K_q scores constructs this output.

For the scalar-envelope claim, suppose

$$\max_{P_q \subseteq C_q} \mathbb{I}[S_q \not\subseteq \mathsf{A}_\gamma(P_q)] \leq \mathbb{I}[V(q) > \gamma] \quad \text{for every } \gamma. \tag{17}$$

Fix q. If $V(q) < Z(q)$, choose a threshold γ strictly between them. Equation 10 makes the left side of Eq. 17 equal to one, while its right side is zero, a contradiction. Hence every scalar witness that dominates the complete action family at every threshold satisfies $V(q) \geq Z(q)$. If equality of the two indicators is required for every threshold and $V(q) > Z(q)$, choosing γ strictly between $Z(q)$ and $V(q)$ makes the left side zero and the right side one. Thus exactness for all thresholds forces $V(q) = Z(q)$.

Finite-frame clause. Condition on the realized proper-training set P and independent fitting randomness. Set $F = \mathcal{F}_0 \setminus P$, $N = |F|$, draw $C \sim \mathrm{SRSWOR}_n(F)$, and write $A = F \setminus C$, $R = N - n$, and $B = \lfloor \alpha R \rfloor + 1$. For ascending rank k, let $t = n - k$ and

$$p_{\mathrm{FP}}(k) = \max_{0 \leq j \leq t} \Pr\{X_j \leq j\}, \qquad X_j \sim \mathrm{Hypergeom}(N, B + j, n). \tag{18}$$

Choose the smallest k with $p_{\mathrm{FP}}(k) \leq \delta$, or $+\infty$ if none exists. This probability controls the complete random complement A, without conditioning on the later test identities.

Fix the proper-training set and scorer as stated, so $F = \mathcal{F}_0 \setminus P$ contains $N = N_0 - |P|$ fixed question-family maxima. The conditioning excludes the calibration draw and the identities of later test and unassigned slices. Let $C \sim \mathrm{SRSWOR}_n(F)$, $A = F \setminus C$, $t = n - k$, and $B = \lfloor \alpha R \rfloor + 1$. Failure of Eq. 14 is the event

$$E = \big\{ |\{q \in A : Z_q > Z_{(k)}\}| \geq B \big\}.$$

Assign every frame unit a fixed unique secondary label u_q and refine the scores lexicographically as $W_q = (Z_q, u_q)$. The primary coordinate of the kth sampled W order statistic is $Z_{(k)}$. Every unit with $Z_q > Z_{(k)}$ also exceeds this refined order statistic, whereas some tied units may additionally exceed it. Thus E is contained in the corresponding failure event for the distinct W ordering.

Rank the N distinct W values from smallest to largest, and let r be the population rank of the kth calibration order statistic. Exactly $t = n - k$ sampled values exceed it, so $N - r - t$ values in A exceed it. The refined event fails precisely when $r \leq N - (B + t)$. Equivalently, if H_t is the fixed set of the $B + t$ largest frame units and $X_t = |C \cap H_t|$, failure is $X_t \leq t$. Because C is an SRSWOR,

$$X_t \sim \mathrm{Hypergeom}(N, B + t, n), \qquad \Pr(E) \leq \Pr(X_t \leq t).$$

For completeness, choose nested sets H_j of the $B + j$ largest units and write $X_j = |C \cap H_j|$. The implication $X_j \le j \Rightarrow X_{j+1} \le j + 1$ shows that $\Pr(X_j \le j)$ is nondecreasing in j. Hence the final display equals the maximum in Eq. 18. All-distinct score populations attain equality, proving worst-case sharpness. Increasing k decreases t, so the smallest admissible k is the least conservative valid calibration rank. If no rank is admissible, $\gamma = +\infty$ makes the strict-exceedance event empty. For multiple preregistered frames sharing one registered real-valued scorer and score scale, replacing each frame-specific threshold by $\gamma = \max_g \gamma_g$ cannot increase any strict-exceedance proportion. If scorers or their numerical scales differ, the same argument instead applies γ_g only within frame g. A union bound over the frame-specific calibration-failure events gives simultaneous confidence $1 - \sum_g \delta_g$, without cross-frame independence; any fixed convex mixture is then bounded by α.

The inherited event is

$$\mathbb{I}[\exists r : h_q^{\mathrm{stop}}(a_{q,r}) = 1] \le B_\gamma(q). \tag{19}$$

For question q and retriever r, write the nested evidence states as $E_{q,r,1} \subseteq \cdots \subseteq E_{q,r,T_{q,r}}$, and let S_q be the annotated support set. Define $h_{q,r}(t) = \mathbb{I}[S_q \not\subseteq E_{q,r,t}]$. Nested prefixes make $h_{q,r}(t)$ non-increasing in t. The registered risk-candidate family contains every prefix that can become a first strict score exceedance. Let $\tau_{q,r}$ be the earliest prefix whose frozen score strictly exceeds γ. When an exceedance exists, the derived policy executes $a_{q,r} \ge \tau_{q,r}$; otherwise it executes the full prefix.

Suppose the derived policy is harmful for some r. The full-prefix sufficiency premise rules out the no-exceedance fallback. Since $\tau_{q,r} \le a_{q,r}$ and harm is non-increasing, $h_{q,r}(a_{q,r}) = 1$ implies $h_{q,r}(\tau_{q,r}) = 1$. The registered proposal is therefore both harmful and above threshold, so $M(q) > \gamma$ and $B_\gamma(q) = 1$. Taking the union over $r \in \mathcal{R}$ proves Eq. 19. A post-admission gate that rejects the cell executes the sufficient full prefix and has zero harm; one that admits the cell retains the preceding inclusion. Thus the gate cannot enlarge the certified harmful-event envelope. The pointwise inclusion holds independently of gate freezing. Prospective utility inference uses a gate frozen before cohort outcomes and a separate F1/EM test.

End-to-end composition. Let $c_{q,r}^{\mathrm{out}}$ be the final action, including fallback, and let $a_{q,r}(c)$ indicate selection by the fixed arbitration rule. Candidate and final-output harms are

$$\begin{aligned} H_{\mathcal{C}}(q) &= \mathbb{I}[\exists r,\ c \in \mathcal{C}_{q,r} : a_{q,r}(c) h_q(c) = 1], \\ H_{\mathrm{end}}(q) &= \mathbb{I}[\exists r : h_q(c_{q,r}^{\mathrm{out}}) = 1]. \end{aligned} \tag{20}$$

Without post-admission transformation, a zero-harm fallback gives $H_{\mathrm{end}} = H_{\mathcal{C}}$. A transformed action instead needs direct containment of its final harm.

Proposition 1 (certified imperfect fallback). Under the conditions of Corollary 1 except its zero-harm fallback assumption, assume $\delta + \eta \le 1$, let $\mathcal{D}$ collect the full registered calibration information, and let $H_F(q)$ indicate selection of a harmful registered fallback. If the harmful-candidate-envelope and fallback certificates respectively give $\Pr(B_\gamma(q) = 1 \mid \mathcal{D}) \le \alpha$ and $\Pr(H_F(q) = 1 \mid \mathcal{D}) \le \beta$ with confidences $1 - \delta$ and $1 - \eta$ for the same frozen policy and deployment distribution, then $\Pr(H_{\mathrm{end}}(q) = 1 \mid \mathcal{D}) \le \min\{1, \alpha + \beta\}$ with confidence at least $1 - \delta - \eta$. Finite-frame composition requires both certificates to target the same random $A = F \setminus C$; a fail-closed zero-harm abstention sets $\beta = 0$.

For the closed-frame route, a harmful direct action must exceed the calibrated threshold and therefore implies $B_\gamma(q) = 1$ (or $Z_q > \gamma$ in the finite-frame regime). A zero-harm fallback cannot add harm, and every permitted post-admission transformation must preserve this pointwise containment. For context cutoff, the nested-execution argument above supplies containment even when the later executed prefix does not itself exceed the threshold. Thus Corollary 1 applies to the final action, not just the initially admitted proposal.

For Proposition 1, remove the zero-harm fallback assumption and let $H_F(q)$ be the event defined there. Every harmful final output is then witnessed either by the harmful-candidate envelope or by a selected harmful fallback, giving $H_{\mathrm{end}}(q) \le \mathbb{I}[B_\gamma(q) = 1 \text{ or } H_F(q) = 1]$ pointwise. A union bound yields the population risk or finite-frame proportion $\min\{1, \alpha + \beta\}$ whenever both certificates hold; in the finite-frame case they must concern the exact same random $A = F \setminus C$. A second union bound over their calibration-failure events gives confidence at least $1 - \delta - \eta$; no independence between the two calibration procedures is required.

A.4 Proof of Corollary 2

Fixed-policy clause. Write $A = \mathbb{I}[\pi(T) \neq \perp]$ and $L = \ell(Q, T, \pi(T))$. Registration gives $L = 0$ on $A = 0$. Since the complete tuples (Q_i, T_i, A_i, L_i) are i.i.d. and π is fixed before the audit, whenever $p_A = \Pr(A = 1) > 0$ the accepted losses have common parameter

$$r = \Pr(L = 1 \mid A = 1) = R_{\mathrm{sel}}(\pi).$$

Condition on the full answer-indicator vector $(A_1, \ldots, A_n) = a$ with $m = \sum_i a_i > 0$. Independence of the original tuples implies that the losses at the m accepted indices remain mutually independent Bernoulli(r) variables. Hence

$$X = \sum_{i=1}^{n} L_i \mid (A_1, \ldots, A_n) = a \sim \mathrm{Binomial}(m, r).$$

The one-sided Clopper–Pearson endpoint in Eq. 9 is the exact inversion of the lower binomial tail, so for every such a,

$$\Pr\{r > U_{\mathrm{CP}}(X, m; \delta) \mid A_1, \ldots, A_n = a\} \leq \delta.$$

If $m = 0$, the convention $U_{\mathrm{CP}} = 1$ makes the event impossible. Averaging the conditional inequalities over every answer-indicator vector, including the random count M, yields

$$\Pr_{\mathrm{audit}}\{R_{\mathrm{sel}}(\pi) > U_{\mathrm{CP}}(X, M; \delta)\} \leq \delta.$$

No part of this conditioning opens the trace: arbitrary dependence among steps, candidate actions, retrievers, and stopping decisions has already been absorbed into the distribution of the single random object T.

Finally,

$$R_{\mathrm{end}}(\pi) = \Pr(A = 1) \Pr(L = 1 \mid A = 1) = p_A R_{\mathrm{sel}}(\pi) \leq R_{\mathrm{sel}}(\pi),$$

where the equality uses $L = 0$ under abstention. This proves Eq. 15. A direct Clopper–Pearson inversion of $X \sim \mathrm{Binomial}(n, R_{\mathrm{end}})$ gives a separate, often tighter marginal endpoint for R_{end}; multiplying the random sample coverage M/n by the selective endpoint is not used as a confidence bound.

For a finite library Π fixed before the audit, apply the preceding result to each $\pi \in \Pi$ at tail error $\delta/|\Pi|$. A union bound makes all intervals valid simultaneously with probability at least $1 - \delta$, so any label-dependent selection among them is covered. If a separate development sample selects π, condition on that sample; the chosen policy is fixed relative to the independent audit and the single-policy result applies.

Simultaneous multi-loss and maximum-coverage clause. For every fixed pair (π, j), the fixed-policy clause applied to ℓ_j at tail error $\delta/(J|\Pi|)$ gives

$$\Pr\{R_{\mathrm{sel},j}(\pi) > U_{\pi j}\} \leq \frac{\delta}{J|\Pi|}.$$

This application is valid even though the $J|\Pi|$ statistics share the same traces: Corollary 2 uses only independence across complete query traces, not independence across policies or loss coordinates. A union bound over all registered pairs therefore gives

$$\Pr\{\exists(\pi, j) : R_{\mathrm{sel},j}(\pi) > U_{\pi j}\} \leq \delta.$$

On the complementary simultaneous event, every pairwise statement holds before any policy is selected. Hence an arbitrary audit-measurable selector, including one that checks every upper endpoint and then compares coverages, cannot invalidate a coordinate. Zero registered loss under abstention separately yields $R_{\mathrm{end},j}(\pi) \leq R_{\mathrm{sel},j}(\pi)$ for every pair.

Let $\mathcal{F} = \{\pi \in \Pi : U_{\pi j} \leq \alpha_j \ \forall j\}$ and choose the first member of $\arg\max_{\pi \in \mathcal{F}} M_\pi$ under a preregistered tie order. When $\mathcal{F} \neq \varnothing$, the preceding simultaneous event certifies every loss limit of the chosen policy, and the definition of the argmax proves that no other certified member has greater audited coverage. When $\mathcal{F} = \varnothing$, the registered procedure returns no answering policy and fails closed. If independent development data select one π before audit, condition on those data and union-bound only its J coordinates, replacing $\delta/(J|\Pi|)$ by δ/J.

Incomplete outcomes. For each policy and loss, suppose the entire answer-indicator vector is known but some answered losses are unobserved. Let X^+ be observed errors plus the number of missing answered losses. The full, possibly unobserved count X satisfies $X \le X^+$ for every realization. Since the exact-binomial upper endpoint is nondecreasing in its error count, the event $R_{\mathrm{sel}} \le U_{\mathrm{CP}}(X, M; \epsilon)$ is contained in $R_{\mathrm{sel}} \le U_{\mathrm{CP}}(X^+, M; \epsilon)$. The same simultaneous guarantee therefore holds for the larger endpoints, for arbitrary loss-dependent missingness. The fixed policies must define the missing outputs and losses even when execution has not materialized them. The resulting endpoint bounds risk under arbitrary loss-dependent missingness.

Pointwise optimality clause. For $x < m$, let $u_x = U_{\mathrm{CP}}(x, m; \delta)$, so the defining binomial-tail equation is

$$\Pr_{u_x}\{X \le x\} = \delta.$$

The endpoint is nondecreasing in x and has exact one-sided coverage by Corollary 2. Suppose another deterministic nondecreasing exact upper bound V were strictly smaller at some count x: $V(x, m; \delta) < u_x$. Choose r strictly between these two values. Monotonicity gives $V(y, m; \delta) \le V(x, m; \delta) < r$ for every $y \le x$, and therefore

$$\Pr_r\{r > V(X, m; \delta)\} \ge \Pr_r\{X \le x\} > \Pr_{u_x}\{X \le x\} = \delta,$$

where the strict inequality follows because the binomial lower tail is strictly decreasing in r for $x < m$. This contradicts exact coverage. At $x = m$, any valid upper endpoint must equal one: if $V(m, m; \delta) < 1$, choosing $r \in (V(m, m; \delta), 1)$ sufficiently close to one makes $\Pr_r(X = m) = r^m > \delta$, again violating coverage. Hence no coordinate of U_{CP} can be lowered within the stated class.

B Risk Protocol and Prospective Confirmation

All splits are keyed by public question identifiers. Proper-training, calibration, development, and test roles are disjoint within each frozen protocol, and BM25, E5, and BGE trajectories for one question remain grouped. These retrieval baselines follow the standard probabilistic relevance framework and the published E5 and C-Pack embedding releases (Robertson & Zaragoza, 2009; Wang et al., 2022; Xiao et al., 2023). The family maximum is computed before threshold selection. Table 6 shows why the calibration unit must match the complete action family.

Table 6: Calibration-unit ablation with the scorer, candidates, and five folds fixed. Q-harm is over 1,000 question clusters; safe early is over 3,000 retriever–question decisions. Question–retriever max pools each correlated retriever path as a unit; ours groups all three paths into one question–family unit. 'Yes' marks a family-valid rule within each OOF policy; over-budget folds report the empirical audit.

Calibration rule	Family-valid?	α	Q-harm (%)	Safe early (%)	Over-budget folds	$+\infty$ thresholds
Candidate pooled	no	5%	13.80	61.70	5/5	0/5
Retriever independent	no	5%	3.20	29.10	0/5	0/15
Retriever Bonferroni	yes	5%	0.00	0.00	0/5	15/15
Question–retriever max (pooled)	no	5%	5.80	42.73	4/5	0/5
Registered-family question max (ours)	yes	5%	1.00	18.10	0/5	0/5
Candidate pooled	no	10%	24.80	77.43	5/5	0/5
Retriever independent	no	10%	11.40	58.87	2/5	0/15
Retriever Bonferroni	yes	10%	0.50	1.73	0/5	0/15
Question–retriever max (pooled)	no	10%	13.30	60.60	4/5	0/5
Registered-family question max (ours)	yes	10%	5.00	40.37	0/5	0/5

Across the seven registered risk budgets, the full retrieval-native scorer has higher point-estimate safe-early coverage than current-score, score-dynamics, and step-only baselines. A 20,000-replicate simultaneous max-t audit separates it from current-score and score-dynamics for $\alpha \geq .03$; the two $\alpha = .02$ lower bounds include zero. Figure 2 reports the point estimates.

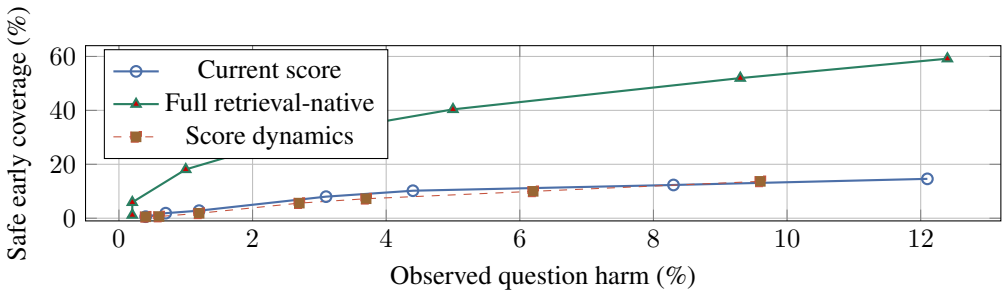


Figure 2: Seven-point OOF frontier over the registered retriever family; the x-axis reports empirical question harm.

B.1 Prospective complete-frame finite-population confirmation

Before allocation or score access, the protocol enumerates the complete official HotpotQA and 2Wiki validation frames, containing 7,405 and 12,576 questions. It freezes the 15-feature scorer, BM25/E5/BGE family, strict threshold comparison, sufficient full fallback, $\alpha = .05$, and total $\delta = .05$ split equally across datasets. Timestamped public randomness selects 2,000 calibration IDs without replacement from each ordered frame. The exact ranks are 1,922 for HotpotQA and 1,921 for 2Wiki; their hypergeometric failure bounds are .02424 and .02326, below the prespecified .025 allocations.

Table 7: Prospective complete-frame confirmation at $\alpha = .05$ and total $\delta = .05$. Envelope and selected-harm counts are question-family unions over BM25, E5, and BGE; early action is measured over question–retriever actions. Caps, thresholds, and nonvacuity gates were frozen before allocation.

Dataset	N	n	R	Envelope / cap	Selected harm	Contain. viol.	Early action
HotpotQA	7,405	2,000	5,405	179 / 270	179 (3.31%)	0	6,468 / 16,215 (39.89%)
2Wiki	12,576	2,000	10,576	398 / 528	398 (3.76%)	0	10,043 / 31,728 (31.65%)

All terminal gates pass. The envelope counts are 179 and 398, below the frozen caps of 270 and 528, and no selected harmful action lies outside its family envelope. The independent replay reconstructs the complete frames, public-randomness allocation, exact ranks, features, scores, BM25 trajectories, and report from receipt-bound E5/BGE trajectories. This certifies the support-omission envelope on the random complements.

C Theorem-Aligned MuSiQue Confirmation

The final protocol allocates 3,000 calibration and 2,000 test questions from the answerable MuSiQue training frame after excluding every ID used by scorer development or earlier diagnostics. Before opening either role, it freezes the deletion scorer, strict comparison, $\alpha = \delta = .05$, Eq. 8 with $P_q = \varnothing$, four comparison contexts, Qwen2.5-7B greedy decoding, 50,000 question-paired bootstrap draws with seed 2026090204, and five terminal gates. The calibration rank is 2,870; its threshold is .908956 and its exact binomial tail is .048846.

Every method receives the same 2,000 questions and 39,992 candidate paragraphs. Full evidence retains all candidates. The proposal-agnostic action retains every item whose deletion score is at most the calibrated threshold. Native Conformal-Embedding uses Qwen3-Embedding-8B revision

b38b5852, the official retrieval instruction, last-nonpadding-token pooling, L2 normalization, and its positive-calibration quantile. The Qwen3 dense comparator is then forced to retain exactly the proposal-agnostic action's cardinality separately on every question. There are zero candidate, decoder, or equal-budget mismatches.

Table 2 reports the terminal values. The proposal-agnostic action omits support on 83/2,000 questions (4.15%; one-sided 95% upper bound 4.96%) and removes 47.40% of candidate evidence (one-sided 95% lower bound 46.70%). Native CCE omits support on 201 questions; the paired native-minus-ours lower endpoint is $+4.35$ points. The exact-budget dense comparator omits support on 847 questions. Against full evidence, proposal-agnostic F1 and EM improve by $+.0527$ and $+.0445$; their one-sided 97.5% lower bounds are $+.0365$ and $+.0280$. All five frozen gates therefore pass.

Generation uses two byte-identical token replicas for each of 7,359 unique states. Verification reconstructs all 8,000 method–question rows and the registered resamples; the Reproducibility Statement identifies the corresponding verification records.

C.1 Factorial native risk–budget–coverage replay

Table 8: **Answer yield under a common support-risk screen.** Secondary best-yield cross-sections of the generated 1,920-point grid. Bounds allocate $.05/(2 \cdot 1920)$ per loss/point. Coverage, correct answers per allocated question (yield), and upper bounds are percentages. Dash: no grid point passes the support screen. Answer-error bounds describe each selected point.

Cap	Scorer	Coverage	EM yield	Support UB	EM-error UB	Prompt M
10	Qwen3 embedding	—	—	—	—	—
10	Learned support	20.55	9.75	4.38	62.84	.457
15	Qwen3 embedding	—	—	—	—	—
15	Learned support	82.10	33.10	4.91	64.72	2.400
20	Qwen3 embedding	100.00	33.55	1.92	70.81	5.091
20	Learned support	100.00	37.80	4.61	66.70	3.210

We reuse the verified scalar scores on all 3,000 calibration and 2,000 test questions. The secondary grid crosses Qwen3 embedding and learned-support scores with native positive-pair and complete-action maximum calibration; it includes 24 nominal risks ($.005, .010, \ldots, .100, .125, .150, .200, .250$) and every hard budget from 1 to 20 paragraphs, for 1,920 points. All thresholds depend only on calibration scores and support labels. A budget wrapper abstains when the threshold-selected set exceeds the cap; it never truncates that set. The question-max embedding arm is an explicit adaptation of the native baseline. The original support-only replay is followed below by complete answer generation for this same registered grid.

Table 3 reports the $.05$ nominal-risk, uncapped anchors. The 6,998 positive pairs give native rank 6,650; the 3,000 question maxima give rank 2,870. On the embedding scorer, the respective thresholds are .757082 and .783047; on the learned scorer they are .821199 and .908956. Consequently the same-scorer comparison isolates which event is calibrated, whereas the same-calibrator comparison measures score quality. Both calibrators sweep the same threshold action family for a fixed scorer, so a fully tuned native threshold can realize the same action as its question-max adaptation.

Within the observed grid, at an empirical selective-omission ceiling of 5%, the learned scorer reaches 46.95% coverage with a 10-paragraph cap and 90.40% with a 15-paragraph cap. No positive-coverage embedding-grid point meets that empirical ceiling under either cap. These support-only cross-sections describe the observed test grid. The complete machine-readable ledger supports independent vector/count reconstruction of all 1,920 points and both original method anchors.

We additionally enumerate every distinct test-score breakpoint: 39,895 for Qwen3 embedding and 39,992 for learned support. This covers every global-threshold retained set on this test for either scorer, including thresholds unavailable in the calibration grid. Both methods receive the same hard-budget rejection rule and unrestricted hindsight threshold selection. Independent binary-search replay checks all 1,597,740 threshold–budget points. At an empirical 5% support-omission ceiling, the learned scorer's oracle coverage is 49.05% under a 10-paragraph cap and 91.25% under a 15-paragraph cap; no positive-coverage native embedding threshold satisfies that ceiling at either cap. At full coverage,

the smallest mean counts are 9.955 and 18.423, respectively. Table 9 gives this cross-section. This descriptive upper envelope isolates score quality within the fixed scorer's global-threshold class on the observed test set.

Table 9: **Exhaustive threshold oracle at empirical support omission $\leq 5\%$.** Both frozen scorers optimize over every threshold breakpoint and use the same hard evidence cap. A dash means that no positive-coverage threshold is feasible. All thresholds are hindsight selections on the opened test. Mean kept is conditional on answering.

	Answer coverage (%)		Mean paragraphs kept	
Hard cap	Embedding	Learned	Embedding	Learned
5	–	7.05	–	4.26
10	–	49.05	–	7.53
15	–	91.25	–	9.40
20	100.00	100.00	18.42	9.96

A separate protocol freezes the entire 1,920-point grid, the original Qwen2.5-7B snapshot, prompt, greedy decoder, and evidence ordering before generation. The 96 distinct scorer–calibrator–risk cells produce 192,000 question–cell joins and 38,152 unique full generation states. Exactly 6,034 states reuse an original answer with identical full input and model/decoder identity; 32,118 use new generation with two identical token replicas. Hard-budget rejection depends only on context size, so one cell answer can be reused across its 20 budget choices. The report covers the complete grid of risks and coverage levels.

An independent implementation rebuilds every join, redecodes saved tokens, recomputes normalized alias-maximized EM/F1 and all budget decisions, and checks 7,680 binomial endpoints. Each point receives both pointwise .025 tails and simultaneous $.05/(1920 \cdot 2)$ tails for its two conditional risks. Table 8 uses the latter to screen support risk, then maximizes correct-answer yield descriptively within each scorer and cap. Its chosen cells are learned positive-pair .005 at cap 10, learned maximum .035 at cap 15, embedding positive-pair .005 at cap 20, and learned maximum .040 at cap 20. Ties in yield prefer lower total prompt cost. Their exact answered/correct counts are 411/195, 1,642/662, 2,000/671, and 2,000/756. The answer-error bounds exceed 25%, illustrating the distinction between support preservation and answer correctness.

At the prespecified .05, cap-20 anchors, six pairwise comparisons use 50,000 paired bootstrap draws each with seed 2026090802. Intervals are descriptive 95% intervals, not policy-selection tests. Under learned scores, maximum minus positive-pair calibration changes support omission by -6.05 points, EM by -2.15 ($[-3.55, -.75]$), and F1 by -2.00 ($[-3.38, -.63]$). Under maximum calibration, learned minus embedding changes EM by $+4.70$ ($[3.10, 6.30]$) and F1 by $+5.43$ ($[3.83, 6.99]$). The corresponding prompt totals are 3,014,036 and 4,838,979, a 37.71% reduction. These comparisons identify the benefit of the scorer and the price of stronger support retention separately.

For a deployed grid point, cost counts each admitted question's complete answer once; hard-budget rejection can occur before generation. Counterfactual generations and duplicate verification used to produce the full grid are additional experimental work, not deployment savings. The original MuSiQue audit uses its one-sided .05 support endpoint (4.96%); the grid uses a more conservative simultaneous endpoint of 6.34% for that same anchor. The counts agree; the multiplicity allocations differ.

C.2 Convex closure of the registered matched answer frontier

The descriptive convex analysis uses a fixed scenario grid and the recorded answer ledger. It includes native embedding (480 points), enhanced embedding (both calibrators, 960 points), and learned scoring (both calibrators, 960 points). Within each group it permits every trace-independent mixture of registered points and abstention. The hard paragraph caps are 5, 10, 15 and 20; mean generation-prompt budgets are 500, 1,000, 1,500 and 2,500 tokens per allocated query; answer-risk limits are $.25, .50, .75$ and support limits are $.05, .10$. Both coverage and correct-answer yield are optimized, giving 576 linear programs. The questions, candidate pool, frozen score vectors, generation model, prompts, evidence ordering and decoder remain unchanged.

Table 10: **Randomized hindsight frontier under equal budgets.** Best correct answers per 100 queries over the frozen finite menu and abstention. All rows impose empirical support risk $\leq 5\%$, answer error $\leq 75\%$, and mean generation-prompt cost $\leq 1{,}500$ tokens per query. Both scorers share candidates and decoder; "enhanced" gives the embedding baseline both calibrators. A zero denotes no positive feasible answer mass. Entries are empirical finite-menu optima.

Hard paragraph cap	Native embedding	Enhanced embedding	Learned
10	0.00	0.00	22.48
15	0.00	0.00	36.05
20	22.65	22.77	38.59

For point i, let c_i, a_i, s_i, t_i be empirical answered mass, answer-error mass, support-error mass, and prompt tokens per allocated query. With weights $w_i \geq 0$, implicit abstention mass $1 - \sum_i w_i$, and common cost cap T, the constraints are

$$\sum_i w_i \leq 1, \quad \sum_i w_i(a_i - \alpha_a c_i) \leq 0, \quad \sum_i w_i(s_i - \alpha_s c_i) \leq 0, \quad \sum_i w_i t_i \leq T.$$

The objectives are $\sum_i w_i c_i$ and $\sum_i w_i(c_i - a_i)$. Selecting a component before a fresh query makes these exact expectations over the empirical query distribution. Components above the common hard paragraph cap are excluded. Prompt cost includes generation for every admitted query; a hard-budget rejection occurs before generation. It excludes scorer/retriever computation and experimental duplicate inference. The matched cost is generator-input tokens.

Independent primal–dual verification reconstructs all 1,920 points from 192,000 recorded answers. All 576 programs satisfy the inequalities and duality-gap tolerance 10^{-8}; the maximum residual and gap are each 2.23×10^{-16} after upward rounding. Thus the finite-menu optimum includes randomized mixtures, not only grid vertices.

At support limit .05, answer limit .75 and cost 1,500, the enhanced embedding optimum at cap 20 has coverage 64.02%, answer error 64.44%, and correct yield 22.77%. The learned optimum has coverage 100%, answer error 61.41%, correct yield 38.59%, and cost 1,439.27 tokens. At cap 15 its correct yield is 36.05%, while the embedding menu has no positive feasible mixture. The stricter answer limit .25 reduces learned correct yield to 0.3472% (coverage 0.4630%) and leaves embedding at zero under the same support constraint. These empirical optima use observed outcomes and cover all trace-independent mixtures of the registered finite thresholds. The supplementary directory `traces/convex_frontier/` contains the results and primal–dual witnesses.

D Open-Corpus Selective Certificate and Transfer

The fixed-pool route and the random-trace route make different reductions. In the official Wikipedia-abstract materialization, Top-20 fallback misses support on 74.505% of HotpotQA and 90.223% of 2Wiki questions; even Top-100 misses on 65.643% and 86.857%. Corollary 2 addresses this failure mode by placing the full query-dependent retrieval computation inside T, permitting the realized 4–17-article actions to vary across questions, and executing $\perp$ when the frozen selector does not meet its answer rule.

The first open-corpus protocol uses an independent 4,000-question calibration gate and two mutually disjoint 10,000-question tests. The frozen threshold answers 2,924 and 2,856 questions, with 100 and 86 support omissions. Their one-sided 95% selective-risk upper bounds are 4.03% and 3.59%; direct all-question bounds are 1.18% and 1.03%. Table 11 reports these earlier support-only audits.

A third disjoint 10,000-question role audits answer correctness as well as support. Development selects threshold .935 from six candidates and freezes a 35% coverage floor, 5% support-risk limit, 50% answer-error ceiling, and total $\delta = .05$ split across the two losses. BM25/E5/BGE retrieval and Qwen2.5-14B greedy decoding are fixed. Of 3,917 answers, 145 omit support and 929 fail exact match; the one-sided 97.5% upper bounds are 4.34% and 25.08%. Mean answered-question token F1 is 84.59%. The audit meets its registered 50% answer-error target; the following audit uses a 25% target.

The six thresholds were registered before this third test. A secondary simultaneous audit uses $.05/12$ per policy–loss pair. At .925, the policy admits 531 more answers than the primary operating point: 4,448 answers in total, with 183 support omissions and 1,053 EM errors. The corresponding upper bounds are 0.0496400194593 and 0.2539540404423. This secondary menu analysis uses the same test cohort. The subsequent strict replication freezes a policy before allocating new questions.

Table 11: **Random-trace HotpotQA certificates on disjoint 10,000-question tests.** The first two rows certify support omission at one-sided 95% confidence. The final row jointly certifies support omission and exact-match answer error at one-sided 97.5% confidence per loss under Corollary 2.

Audit	Answered	Coverage (%)	Support errors	Support upper (%)	Answer EM errors	Answer upper (%)	Gate
Support test 1	2,924	29.24	100	4.03	–	–	pass
Support test 2	2,856	28.56	86	3.59	–	–	pass
Joint test	3,917	39.17	145	4.34	929	25.08	pass

D.1 Prospective 20,000-question strict replication

Before allocation, the protocol fixes threshold .925, a 5% support-omission ceiling, a 25% normalized answer-error ceiling, and at least 40% empirical coverage. The two losses each receive tail error .025. Allocation excludes the 66,822-ID union of previously opened roles, leaving 31,030 eligible HotpotQA train-frame IDs; a fixed-salt SHA-256 ordering selects 20,000. Full reconstruction verifies zero overlap. The binomial population interpretation of this deterministic archive allocation uses the complete-trace i.i.d. assumption in Corollary 2.

BM25, E5, and BGE each produce all 20,000 rankings. The unchanged policy admits 8,896 distinct generation states. Qwen2.5-14B uses the frozen prompts and greedy 16-token decoding rule. The first row of Table 13 has 352 support omissions (3.96%) and 2,224 normalized exact-match errors (25.00%), with respective upper bounds .0438291008437 and .2591352882860. Coverage is .4448, mean answered token F1 is .840061116420, and EM is .75. The stricter joint gate fails on answer error; no threshold fallback, optional stopping, or post-test policy selection is applied to this audit.

Answer EM/F1 use the frozen project scorer: Unicode NFKD normalization, lowercasing, word-boundary removal of "a," "an," and "the," replacement of non-`[a-z0-9]` runs with spaces, and whitespace collapse. EM is normalized-string equality and F1 uses token-multiset overlap, each maximized over registered gold aliases; the answer-error loss is $1 - \mathrm{EM}$.

Verification reproduces every decision, recorded answer, loss, and confidence endpoint.

D.2 Fresh 10,000-question post-answer audit

The 20,000 opened questions are used only for development, split into 10,000 fitting, 5,000 selection, and 5,000 validation questions for a 225-candidate library. The selected rule returns 8,507 development answers, with 1,842 EM errors and 280 support omissions. These development counts do not enter the fresh audit below.

The frozen policy preserves the .925 support-score threshold, Qwen2.5-14B revision `cf98f3b3`, BM25/E5/BGE retrieval, prompts, and greedy 16-token generation limit. It normalizes the generated answer with the unchanged scorer. It rejects an answer containing any registered uncertainty string: "unknown," "cannot," "not provided," "not enough," or "not mentioned." Otherwise it returns a nonempty answer only if the normalized answer is a contiguous whole-word sequence in at least one supplied evidence text, or is exactly "yes" or "no." The gate uses no gold alias, support annotation, or question ID. Evidence-text occurrence enters answer selection; a separate audit measures the complete modified policy's risk.

The scientific protocol fixes this single policy before test allocation, with 10,000 questions, at least 40% empirical coverage, support-omission upper bound at most 5%, and EM-error upper bound at most 25%. Each loss receives tail error .025, giving joint confidence 95% under Corollary 2's

sampling assumption. Excluding the 86,822-ID historical union leaves 11,030 eligible HotpotQA training IDs. A fixed-salt SHA-256 ordering allocates 10,000; independent reconstruction verifies their identities and zero prior-role overlap. This role is question-ID-disjoint within the same archive and uses Corollary 2's sampling assumption. The protocol forbids optional stopping, threshold fallback, test replacement, and policy selection after test outcomes.

All three retrievers complete 10,000 rankings. The base gate creates 4,451 unique generation states; the post-answer gate abstains on 197 and returns 4,254. Among returned answers, 883 are EM errors (20.7569%) and 153 omit support (3.5966%). Separate binomial inversion gives upper bounds .2200731436288431 and .0420071527949532. All three registered gates pass at 42.54% coverage. Mean answered token F1 is .871148599142081 and EM is .792430653502586. These estimates describe the frozen policy on its fresh audit cohort.

Cost includes all 5,571,108 prompt tokens and 21,974 generated tokens, including the 197 rejected generations; generation takes 3,671.86 seconds. The verification checks both the base and post-answer decisions, as well as the recorded inputs and losses.

D.3 Paired attribution of the answer-or-abstain rule

This secondary analysis uses the same 10,000 allocated questions, 4,451 generated answers, supplied evidence, aliases, and support labels as the frozen policy. Its analysis design fixes four arms before the new aggregation: the base .925 selector; base plus uncertainty filtering alone; base plus evidence-text occurrence alone, with the same yes/no exception; and the unchanged complete gate. All four fixed arms reuse the same 4,451 generations and their recorded costs in a paired secondary analysis.

Table 12: **Same-output HotpotQA component ablation.** All arms use the same 10,000 questions and 4,451 generations. UB columns use the original .025 tail convention as a secondary reference. The complete gate's prospective result remains unchanged.

Post-answer component	Returned	Coverage	EM errors	EM-error UB	Support UB
None	4,451	44.51%	1,078	25.51%	4.74%
Uncertainty only	4,450	44.50%	1,077	25.49%	4.71%
Evidence occurrence only	4,254	42.54%	883	22.01%	4.20%
Complete frozen gate	4,254	42.54%	883	22.01%	4.20%

The unfiltered answers contain 1,078 EM errors; the gate removes 195 and returns 883. It retains 3,371 of the original 3,373 correct answers. Thus it removes 18.09% of base EM errors while losing 0.059% of base correct answers, at a 1.97-point coverage cost. Among the 197 rejections, 195 are EM errors (98.98%). Under uniform fixed-cardinality rejection of 197 out of these same 4,451 answers, the number of rejected errors is hypergeometric with expectation 47.712; the observed enrichment is not explained by the smaller answered set alone. The returned-answer error rate decreases from 24.2193% to 20.7569%, whereas random rejection has expected conditional error 24.2193% at exactly the same coverage.

For paired uncertainty estimates, we compare gates on the same 10,000 questions and resample this entire cohort, retaining the 5,549 base abstentions. An equivalent multinomial representation has five categories: kept correct, kept erroneous, rejected correct, rejected erroneous, and base abstention. Fifty thousand draws per loss, with seeds 2026090801 and 2026090802, give descriptive two-sided 97.5% intervals. The full-minus-base conditional EM-error difference is -3.4623 points, interval $[-4.0225, -2.9262]$; the support-omission difference is -0.5148 points, interval $[-0.7960, -0.2618]$. The gate rejects 30 support-omitting answers versus 8.100 expected under matched random rejection.

Uncertainty filtering rejects only one answer, already rejected by the occurrence rule. Occurrence alone therefore matches the complete gate on this cohort, as Table 12 shows. For a more conservative secondary risk reference, allocating $.05/8$ across four arms and two losses gives the complete gate upper bounds 22.35% EM error and 4.37% support omission, below the numerical targets. This secondary four-arm reference complements the prospective two-loss certificate.

D.4 FRESH FIXED-CORPUS 2WIKI STRICT CERTIFICATE

The 2Wiki policy is developed on the target corpus. Its candidate factory retrieves from the fixed native 2Wiki passage corpus and prioritizes title-linked evidence, returning at most 20 passages. Two ridge-logistic models predict answer error and support omission from 31 features: question and answer length and type, answer occurrence and lexical overlap with evidence, retrieval-link counts, and generated-token log-probability summaries. Neither question IDs, gold answers, support annotations, nor role indicators enter inference features. Standardization and model coefficients are fitted on 3,000 development questions. A separate 2,000-question selection role compares 240 candidates. The selected candidate uses probability cutoffs .35 for answer error and .075 for support omission; both must pass to return the generated answer. Its selection-role counts are 846 answers, 82 EM errors, and 27 support omissions.

Before fresh allocation, the scientific protocol freezes both scorers, thresholds, evidence ordering, Qwen2.5-14B-Instruct revision `cf98f3b3`, prompts, and greedy decoding with seed zero and at most 16 new tokens. The complete action includes generation followed by the joint selector. Generation cost includes answers subsequently rejected by the selector. The confirmation generates every distinct test question, including those ultimately rejected, without truncating its registered evidence prompt.

The ID-only historical inventory excludes 43,456 previously used questions from the 167,454-question native training frame, leaving 123,998 eligible IDs. The protocol is frozen before a single OS-entropy seed is published. It draws exactly 10,000 IDs independently and uniformly with replacement. The resulting 9,612 distinct IDs have no historical-role overlap. Repeated IDs retain their draw multiplicity in every count; deterministic generation is computed once per distinct context. The binomial sample size is 10,000 independent draws, with 9,612 distinct question IDs. Conditional on the fixed corpus and deterministic policy, random draws induce i.i.d. answer and loss indicators even when an ID repeats. The estimand is the uniform distribution over these historically unused native-frame questions on the fixed corpus.

Let $N = 10{,}000$, M be answered draws, and X_a, X_s be answered draws with EM error and support omission. The protocol requires

$$U_{\mathrm{CP}}(X_a, M; \eta) \le .25, \qquad U_{\mathrm{CP}}(X_s, M; \eta) \le .05, \qquad L_{\mathrm{CP}}(M, N; \eta) \ge .40, \quad \eta = .05/3. \tag{21}$$

For $M > 0$, $L_{\mathrm{CP}}(M, N; \eta)$ is the η quantile of $\mathrm{Beta}(M, N - M + 1)$; it is zero for $M = 0$. A union bound gives simultaneous confidence at least 95% for the two conditional risks and unconditional coverage. This coverage requirement is stronger than a floor on observed coverage alone. The policy is chosen only on separate development data, so the fresh audit does not pay a 240-candidate selection factor. Fixed sample size, no intermediate outcome inspection, no optional stopping, and no post-test threshold change are part of the protocol.

The terminal counts are $M = 4{,}274$, $X_a = 349$, and $X_s = 145$. Independent exact-binomial inversion gives

$$U_a = .0910216072361, \qquad U_s = .0403157355801, \qquad L_{\mathrm{coverage}} = .4168484806542.$$

All three targets pass; observed coverage is $.4274$. These outcomes describe the policy on its allocated 2Wiki audit cohort.

Table 13: **Strict whole-policy audits.** Row 1 reports the HotpotQA 20k audit. Rows 2–3 compare base and post-answer gates on shared outputs from a fresh, ID-disjoint 10k cohort; † marks the paired base-gate reference. HotpotQA uses .025 per risk bound and empirical coverage. The 2Wiki policy uses $.05/3$ per endpoint, including a coverage lower bound (LB); its empirical coverage is 42.74%.

Policy / audit size	Answers	Coverage	Support UB	EM-error UB	Joint gate
Base .925 / 20,000	8,896	44.48%	4.38%	25.91%	Fail
Base, same 10,000†	4,451	44.51%	4.74%	25.51%	Fail†
Post-answer gate / 10,000	4,254	42.54%	4.20%	22.01%	Pass
2Wiki / 10,000 draws	4,274	41.68% LB	4.03%	9.10%	Pass
Registered requirement	–	$\ge 40\%$	$\le 5\%$	$\le 25\%$	All three

Verification checks historical exclusions, allocation, both token and log-probability replicas, selector features, and all draw-weighted losses and bounds. Table 13 collects the completed strict-risk audits.

D.5 SYMMETRIC NATIVE-CORPUS DEVELOPMENT COMPARISON

To separate candidate construction from selective answering, we replay the already opened 3,000 fitting and 2,000 selection questions for both native-corpus BM25@20 and title-linked@20. Both use the same fixed corpus, Qwen2.5-14B prompt and greedy 16-token decoder, the same 31 deployment-visible features, and separately fitted answer-error and support-omission ridge-logistic models. Each receives ridge strengths 1 and 10, ten answer cutoffs, and twelve support cutoffs: 240 candidates per retriever. Standardization and coefficients use only the fitting role. Among selection-role candidates with pointwise .025-tail upper bounds below 25% answer error and 5% support omission, the comparison selects maximum coverage, breaking ties by smaller answer bound, support bound, then candidate index. All 480 candidates are retained. These pointwise selection bounds serve as development diagnostics; the subsequent frozen audit supplies the simultaneous certificate.

Table 14: **Same selector capacity and development budget for both retrievers.** Opened 2,000-question selection role; all percentages refer to this role. Error counts divide by answered questions; coverage and correct yield divide by 2,000. UBs are pointwise development diagnostics. Joint rows are selected from 240 candidates per retriever.

Retriever	Gate	Answered	Coverage	Correct yield	Answer UB	Support UB
BM25@20	None	2,000	100.00	51.15	51.07	67.44
BM25@20	Occurrence	1,567	78.35	49.55	39.20	58.82
BM25@20	Joint	236	11.80	11.15	9.24	4.87
Linked@20	None	2,000	100.00	73.35	28.65	11.61
Linked@20	Occurrence	1,910	95.50	73.35	25.15	10.49
Linked@20	Joint	858	42.90	38.35	12.86	4.82

Both candidate factories retrieve at most 20 passages, but their token costs differ because passage lengths differ. The linked arm spends 4,946,037 prompt tokens on all 2,000 generated answers, versus BM25's 3,895,277; costs include every rejected answer. After joint selection it returns 767 correct answers versus 223, giving 6,449 versus 17,468 prompt tokens per correct answer. This ratio measures generator-input tokens per correct returned answer. Generated-token totals are 9,767 and 10,462, respectively; duplicate verification inference is additional experimental cost.

Neither retriever has an answer-only candidate satisfying both diagnostic risk ceilings. The occurrence-only gate also fails: support upper bounds are 58.82% for BM25 and 10.49% for linked retrieval. Ungated linked retrieval has 73.35% EM but a 10.20% observed support-omission rate; selecting answers and preserving support are distinct objectives. The best linked comparison candidate uses ridge 10 and cutoffs $.40/.075$. The earlier strict-certificate policy uses ridge 1 and cutoffs $.35/.075$. This development comparison uses its separate development cohort.

The development replay reconstructs all 310,000 feature values and 960,000 candidate decisions and checks all 968 binomial endpoints. These selection-role results remain diagnostics, separate from the fresh fixed-policy audit below.

D.6 FRESH ROBUST-SELECTION CONFIRMATION

The six-policy audit excludes all previously allocated or opened IDs. Its final policy uses 19,432 opened development IDs with three repetitions of five-fold ID-isolated cross-fitting. Repeated folds provide model-selection diagnostics over the same development IDs. Separate reconstruction checks 932,736 out-of-fold probabilities and 19,121,088 gate decisions. The selected linked policy has 31 base features, 70 interactions, ridge penalty 100, and answer/support cutoffs $.40/.075$. It is refitted on the opened development pool and frozen before allocation. BM25 retains its frozen baseline joint gate. The six-policy test compares the complete deployed systems, each with its fitted coefficients.

The audit draws 20,000 question IDs independently with replacement from the unused native-corpus population, giving 18,106 distinct questions and 36,207 distinct generation contexts across the two retrievers. Repeated IDs retain their draw multiplicity in the binomial audit. The complete configuration fixes Qwen2.5-14B, the same prompt, Top-20 passage budget, greedy seed 0 and a 16-token output cap. Each generation context has two strictly matching saved token and log-probability replicas. Independent terminal reconstruction checks the exclusion inventory, allocation, retrieval, generation records, six-policy decisions, 18 binomial endpoints, and six 50,000-draw paired-bootstrap comparisons. Verification replays the recorded outputs.

Each policy receives an answer-risk upper bound, support-risk upper bound, and coverage lower bound at tail probability $.025/18$. The six-policy event has confidence at least 97.5% for this audit. The linked joint gate returns 8,819 answered draws, with 808 EM errors and 294 support omissions. Its endpoints are $U_a = .101160356010$, $U_s = .039453365184$, and $L_c = .430443074961$. BM25 joint gating meets both risk limits but returns only 2,161 draws, giving $L_c = .101582829204$. Occurrence gating on linked retrieval returns 19,051 draws, with $U_a = .233449333883$, $U_s = .100605007384$, and $L_c = .947883858035$; it fails the support limit. Table 4 reports every policy.

The linked-joint minus BM25-joint coverage difference is $+.3329$, descriptive paired 95% interval $[.3263, .3395]$; correct-answer yield differs by $+.2986$, interval $[.29215, .30510]$. Linked joint gating sacrifices $.5116$ coverage and $.33835$ correct-answer yield relative to linked occurrence gating. This tradeoff motivates the certificate-derived extension below. All 48,755,290 linked prompt tokens and 96,800 generated tokens count toward system cost, including rejected answers. BM25 uses 38,787,773 prompt tokens and 102,780 generated tokens.

D.7 Retrieval–compression decomposition and a selector-uniform ceiling

Proof of Theorem 1(iv). Fix one trace, its possibly incomplete candidate set C, support S, scores and finite threshold γ. If $B = \mathbb{I}[S \not\subseteq C] = 1$, every retained output inside C omits support, independently of proposal or calibration. If $B = 0$, Theorem 1(i) applies inside C: the empty proposal attains the maximum harmful event $\mathbb{I}[Z_C > \gamma]$. These cases are disjoint, proving Eq. 12. For any retained $A \subseteq C$, the realized loss also decomposes exactly as

$$\mathbb{I}[S \not\subseteq A] = B + (1 - B)\mathbb{I}[S \not\subseteq A].$$

The second term is compression-induced omission conditional on a retrieval-complete trace. One may represent the finite-threshold envelope by $+\infty$ when $B = 1$ and Z_C otherwise. If calibration cannot produce a finite admissible threshold, open retrieval must abstain; restoring the incomplete candidate set does not make it support-complete.

For an arbitrary answer indicator G, including randomized and query-dependent selection, let $c = \Pr(G = 1)$ and $r_s = \Pr(S \not\subseteq A \mid G = 1)$. The event $\{G = 1, S \subseteq A\}$ is contained in $\{B = 0\}$, so $c(1 - r_s) \leq 1 - b$, where $b = \Pr(B = 1)$. This proves the coverage ceiling. It is sharp when an oracle returns the full candidate set on every complete trace and on a fraction of incomplete traces. If $b \leq \alpha_s$, it answers everywhere. Otherwise, it returns extra incomplete mass $\alpha_s(1 - b)/(1 - \alpha_s)$, attaining $c = (1 - b)/(1 - \alpha_s)$ and risk α_s. No realizable selector can exceed this oracle; attaining it may require unavailable completeness labels.

This is the same support-loss envelope as the fixed-frame authorizer, with the irretrievable event made explicit. Answer correctness has its own audit loss. Augmenting C through additional retrieval defines a new completeness event.

The completed 2Wiki pipelines pass all selected Top-20 passages into generation. Their pre-gate support omission is therefore exactly B, using the six-policy audit's normalized supporting-title criterion, with no post-retrieval compression. We retrieve the hash-bound allocation and 36,212 retriever–question joins and independently replay all 20,000 draws per retriever. BM25 has 13,082 missing-support draws; linked retrieval has 2,154. Repeated question IDs retain their original draw multiplicity. The observed missing fractions are $.6541$ and $.1077$; the original joint gates retain respectively 47 and 294 missing-support answers. They reject 13,035/1,860 incomplete draws and 4,804/9,321 complete draws, respectively. These counts separate candidate availability from the selector's operating point.

A secondary audit of the two fixed retrieval policies allocates $.025$ to each lower endpoint, giving simultaneous confidence at least 95%. For m misses in $n = 20{,}000$ draws, the lower bound is the

solution of $\Pr_{b_L}\{\text{Binomial}(n, b_L) \geq m\} = .025$. Direct tail bisection and beta-quantile inversion agree within 10^{-12}. The bounds are

$$b_{L,\text{BM25}} = .647460767942, \qquad b_{L,\text{linked}} = .103435927389.$$

Equation 13 then gives 37.1094% and 94.3752% upper bounds on coverage for *every* downstream subset selector with support risk at most 5%. Thus the 40% target is impossible for this BM25 candidate factory, even with ideal gate tuning. The separately frozen linked gate provides the constructive certificate. Combining both retrieval ceilings with the six-policy 97.5% event gives at least 92.5% confidence by a union bound. The main-text comparison uses only the BM25 ceiling and attains joint 95% confidence, as derived below. The population is the six-policy audit's uniform unused-ID distribution on the fixed native corpus.

Exhaustive tests check all 5,216 frame/support/order/threshold cases on a four-item universe and all 55,712 proposals, including incomplete frames. Additional tests check tie restoration and sharpness of the ceiling. The saved allocation and join hashes match the preparation receipt.

D.8 Joint strict-risk dominance over every same-candidate gate

The ceiling supplies a comparison against a class larger than any fitted gate menu. Fix a candidate factory C_0 and a population. Suppose $b_0 \geq b_L$ holds on an event E_0 of probability at least $1 - \delta_0$, and an alternative policy has coverage $c_1 \geq L_c$, answer risk $r_a \leq U_a$, and support risk $r_s \leq U_s$ on an event E_1 of probability at least $1 - \delta_1$. On their intersection, every gate returning a subset of C_0 with support risk at most α_s satisfies

$$c_0 \leq U_C := \min\{1, (1 - b_L)/(1 - \alpha_s)\}, \qquad y_0 \leq U_C,$$

where y_0 is its correct-answer yield. The second inequality only uses $y_0 \leq c_0$: it remains valid even for a perfect answer decoder. Meanwhile, $y_1 = c_1(1 - r_a) \geq L_c(1 - U_a)$. Thus $L_c > U_C$ certifies a uniform coverage separation, and $L_c(1 - U_a) > U_C$ certifies a uniform correct-yield separation. Their joint confidence is at least $1 - \delta_0 - \delta_1$, without independence or a gate-count correction. The pointwise inclusion behind Theorem 1(iv), rather than enumeration, makes this comparison uniform over randomized, query-dependent, or oracle gates.

For this comparison we allocate $\delta_0 = .025$ to the BM25 lower endpoint. The six-policy event has $\delta_1 = .025$ and already covers the analytic mixture in Appendix D.10. Hence all entries of Table 5 hold jointly with at least 95% confidence. At $\alpha_s = .05$, $U_C = .371093928482$. Frozen linked joint gating gives $y_1 \geq .386899300256$, so its coverage and correct-yield advantages over every admissible BM25 subset gate are at least .059349146479 and .015805371774. The mixture's tighter componentwise bound is $y_1 \geq \sum_i w_i L_i(1 - U_{ia}) = .414075436458$, raising the respective gaps to .100744409125 and .042981507976. Both alternatives also have answer risk below .25.

The supplementary dominance verifier checks the six-policy audit bindings, inverts the binomial tail, and reconstructs these inequalities from the frozen 20,000-draw audit. The comparison uses the original BM25 candidates and supporting-title labels on the six-policy audit's unused-ID population. Each candidate factory and corpus defines its own completeness event.

D.9 Paired online-system cost replay

We replay the two frozen joint-gate systems on 32 previously audited IDs selected by lowest SHA-256 rank under a fixed salt, without consulting outcomes. Each question receives three repetitions per system (192 timed runs); a hash assigns AB/BA order, and both systems receive one untimed warmup. Corpus, retrieval, features, prompts, Qwen2.5-14B weights, fp16 execution, greedy decoding, and the two-GPU configuration are fixed; timing is synchronized. Online time includes retrieval/features, tokenization, generation, log-probabilities, and gating, but excludes model/corpus loading and offline construction or training. All timed runs reproduce the recorded tokens and decisions.

Mean query time is 1.2171 seconds for BM25 and 1.4460 for linked retrieval. Retrieval/features account for .10834/.10830 seconds, prompt processing plus generation for 1.10803/1.33670, and gating for .00078/.00100, respectively. A question-paired bootstrap (50,000 draws; seed 20260911) gives a descriptive 95% interval of $[.1155, .3615]$ seconds for the latency difference and $[1.0933, 1.3022]$ for the mean-time ratio (1.1880). Thus, the certified coverage gain accompanies an observed 18.8% increase in warm-service latency.

D.10 USING A SIMULTANEOUS MENU'S RESIDUAL RISK BUDGET

Randomized selection is standard in reject-option classification (Franc et al., 2023). We extend the complete-action audit's simultaneous bounds to trace-independent mixtures, keeping component scores, losses, and confidence allocations fixed.

On the joint event, let $c_i \geq L_i$ be component i's answer probability and $r_{ij} \leq U_{ij}$ its selective risk under loss j. Choose weights $w_i \geq 0$, $\sum_i w_i = 1$, even after observing the audit; for each fresh trace, draw $I \sim w$ independently of that trace and execute I. Coverage and selective risk are

$$c(w) = \sum_i w_i c_i, \qquad r_j(w) = \frac{\sum_i w_i c_i r_{ij}}{\sum_i w_i c_i},$$

when the denominator is positive. Conditional risk is weighted by answered mass $w_i c_i$, not by arm probability w_i alone. For target α_j, define

$$b_{ij} = \begin{cases} L_i(U_{ij} - \alpha_j), & U_{ij} \leq \alpha_j, \\ U_{ij} - \alpha_j, & U_{ij} > \alpha_j. \end{cases}$$

If $\sum_i w_i b_{ij} \leq 0$ for each loss, then on the same event

$$c(w) \geq \sum_i w_i L_i, \qquad r_j(w) \leq \alpha_j. \tag{22}$$

Indeed, $\sum_i w_i c_i (r_{ij} - \alpha_j) \leq \sum_i w_i c_i (U_{ij} - \alpha_j) \leq \sum_i w_i b_{ij}$, using $c_i \geq L_i$ for negative coefficients and $c_i \leq 1$ for positive ones. The bound holds simultaneously for all w, so audit-dependent selection adds no union-bound penalty; the resulting policy uses fixed components and a trace-independent mixture coin. Maximizing $\sum_i w_i L_i$ subject to the inequalities and simplex constraints is a linear program.

We choose the linked occurrence gate with probability $2/25$ and its frozen joint gate with probability $23/25$. The two rules consume the same generated answer; choosing a gate requires no additional model inference. The ratio above is bounded over $c_i \in [L_i, 1]$. With positive denominator, each coordinate is monotone or constant while the others are fixed, so a maximum occurs at a rectangle vertex. Exact rational evaluation of all four vertices for each of two risks yields

$$L_c = .471838337607, \qquad U_a = .123393441334, \qquad U_s = .0497307873894.$$

Thus the coverage floor improves by 4.1395 percentage points, while both original risk targets remain met on the same simultaneous 97.5% event. Correct-answer yield has lower bound $\sum_i w_i L_i (1 - U_{ia}) = .414075436458$ on that event.

On the recorded traces, expected coverage/yield is 48.1878%/42.7618%. Generation cost matches linked joint gating; cost per correct answer is 5,701 versus 6,086 tokens, including rejections. Deployment uses fresh trace-independent coins on the six-policy audit's uniform unused-ID population and fixed native corpus. Independent verification recomputes all 18 endpoints, eight risk corners, and coverage, yield, and cost identities.