TRIDENT: Selection-Conditional Facet Tests and Episode-Frozen Min-Cost Cover for Budgeted RAG

Anonymous ACL submission

Abstract

Budgeted RAG is a decision problem: under a hard evidence cap, which passages you keep determines both accuracy and what you can credibly claim about the evidence at query time. We introduce TRIDENT, a framework that mines auditable reasoning facets, tests facet support with a calibrated verifier, and selects evidence under an explicit token budget. In the Safe-Cover regime, we freeze the retrieval pipeline into a replayable episode, map verifier scores to selection-conditional conformal p-values under a logged contract, and apply per-query multiple-testing control to yield facet-support certificates—or return a machine-checkable abstention with a reason code. In Pareto-Knapsack, we drop per-query guarantees and optimize a quality–cost frontier for throughput. On HotpotQA at a 500-token evidence cap, TRIDENT Pareto-500 improves EM/F1 from 30.81/39.61 to 45.30/58.22 (+47% relative), while using 3% fewer evidence tokens and 5% lower latency than naive top-k truncation. These results show that under tight budgets, selection rigor and query-time evidence accountability matter as much as retrieval strength.

1 Introduction

Retrieval-augmented generation can improve factuality by grounding a model in retrieved text; however, under strict context limits and latency targets, the bottleneck shifts: the question is no longer whether relevant information can be found, but which evidence is worth retaining under a hard cap. Retrieving more passages can raise accuracy, yet it directly increases token consumption and end-to-end latency while enlarging the surface area for redundant or spurious citations (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2020). Under production constraints—such as hard context limits, SLOs, and GPU contention—the system must do more than rank documents; it must also decide what to keep and when to stop.

Despite progress in adaptive pipelines, such as multi-step retrieval and retrieve–reflect control (Guu et al., 2020; Asai et al., 2023), budgeted evidence selection is still commonly implemented as top-k concatenation plus truncation. Under a hard cap, this policy is brittle for multi-hop questions: truncation can drop a required bridge passage while retaining redundant or merely related text. The opportunity cost is substantial. Even with the retrieval stack held fixed, selection alone can dramatically change outcomes: at an approximate 500-token evidence cap on HotpotQA, our Pareto mode improves EM/F1 from 30.81/39.61 to 45.30/58.22.

A second pain point is query-time accountability. Many RAG systems output citations, but they provide little quantitative statement that the selected passages satisfy the intermediate requirements that a query depends on. Under tight budgets, this matters operationally: a system should sometimes answer, but it should also sometimes refuse—and that refusal should be an auditable decision derived from a rigorous contract, not a post-hoc narrative.

We propose TRIDENT, a budgeted evidence-selection framework that makes the choice of evidence explicit under strict caps. TRIDENT utilizes typed reasoning facets to structure what must be supported, and a verifier signal to determine what evidence is worth paying for. It operates in two regimes: a *Pareto* mode that optimizes the quality–cost frontier under a hard evidence cap, and a *Safe-Cover* mode that emits machine-checkable evidence receipts, or cleanly abstains when the audit contract cannot be satisfied. The end-to-end design is summarized in Figure 1.

Contributions.

- We formalize budgeted RAG as facet-level evidence selection under a hard context budget, with an explicit separation between certifying support and producing an answer, and with

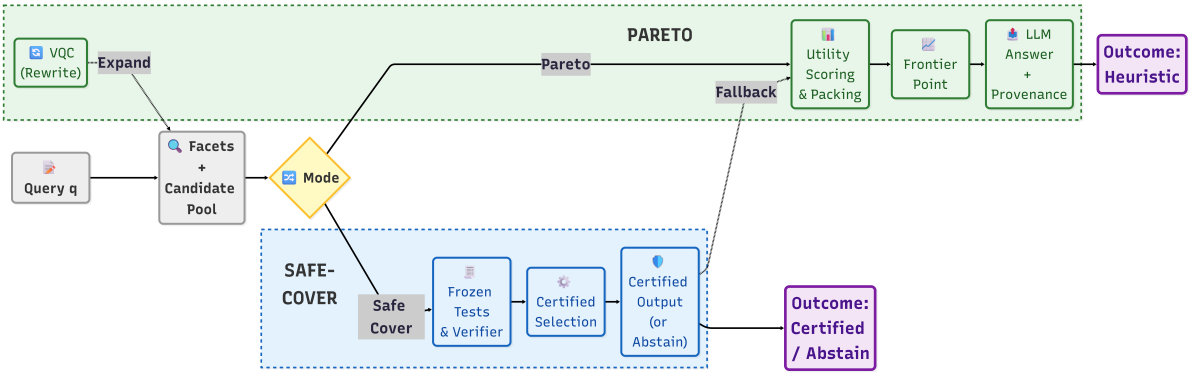

Figure 1: TRIDENT overview. Facets define verifiable requirements, and selection packs evidence under a hard token budget. Safe-Cover certifies facet support or abstains while Pareto-Knapsack heuristically trace a quality–cost frontier without per-query validity claims.

standardized abstention outcomes for evaluation integrity.

- We introduce Safe-Cover, a certified regime that maps verifier scores to selection-conditional conformal p-values, enabling auditable greedy minimum-cost cover or a replay-valid infeasibility certificate under a query-level error budget.

- We evaluate on HotpotQA, 2WikiMultiHopQA, and MuSiQue, reporting EM/F1 together with evidence tokens and latency, with diagnostics that attribute performance to coverage, calibration, and verifier behavior rather than generic failure labels.

2 Related Work

Structured retrieval and reasoning granularity. Multi-hop QA has pushed the field beyond simple independent top-k passage selection. This evolution began with stronger dense retrievers (DPR; Karpukhin et al., 2020) and multi-hop retrieval variants (e.g., MDR; Xiong et al., 2021; BeamDR; Zhao et al., 2021), and has recently matured into structured approaches that organize evidence as graphs, propositions, or memory-like indices (e.g., HippoRAG; Gutiérrez et al., 2025; PropRAG; Wang and Han, 2025). These methods enhance candidate discovery and chaining by modifying indexing granularity or retrieval structure, primarily addressing the recall problem. They do not by themselves address the accounting problem: defining a query-time, machine-checkable statement of which intermediate reasoning requirements are satisfied under a strict evidence budget.

Verifiable selection and efficient RAG. A broad line of reliability work focuses on verifying or repairing model outputs after generation, using self-critique loops (e.g., SELF-RAG; Asai et al., 2023; Corrective RAG; Yan et al., 2024) and attribution frameworks (e.g., FActScore; Min et al., 2023; Li et al., 2024). In parallel, systems-oriented work targets deployment constraints directly, reducing time-to-first-token through caching (Lu et al., 2025), shrinking context via selective augmentation (Mao et al., 2025), or filtering evidence based on estimated utility (Wang et al., 2025). Recent advances in context selection reconceptualize quality assessment as a data valuation problem, measuring each context's marginal contribution through influence-based metrics that capture query-aware relevance, list-aware uniqueness, and generator-aware alignment (Deng et al., 2025). A remaining gap is ex-ante evidence verification: validating intermediate support before generation to avoid spending tokens on low-value or non-entailing evidence, while making the selection decision auditable at query time rather than defended retroactively.

Statistical guarantees in dynamic retrieval. Conformal prediction offers rigorous tools for calibration and uncertainty quantification (Angelopoulos and Bates, 2022), but applying such guarantees in retrieval settings is a subtle task. The policy itself induces the hypotheses being tested: retrieval and shortlisting determine which (passage, requirement) pairs become candidates. Agentic loops that rewrite queries or expand pools can shift the data distribution mid-flight, complicating any per-query validity claim unless the selection mechanism is explicitly controlled and logged. This motivates the "frozen episode" approach, which

involves certificate-style guarantees that are conditional on a locked selection contract (retriever snapshot, shortlist policy, binning), thereby explicitly separating certified regimes from adaptive optimization regimes where no per-query statistical validity is claimed.

3 Framework

We frame retrieval-augmented answering as budgeted, auditable evidence selection with an explicit separation between (i) certifying that evidence supports query requirements and (ii) producing an answer. Given a query q, the system extracts a finite set of intermediate requirements (facets) $F(q)$ and selects a small set of passages S under a hard evidence budget. The framework supports two serving regimes: a certified regime that emits machine-checkable facet-support certificates under a replayable selection contract, and a frontier regime that optimizes quality–cost operating points without per-query statistical claims. In both regimes, abstention is a first-class outcome: rather than hallucinating an answer, the system returns a fixed output token accompanied by an auditable reason code. Figure 1 summarizes the end-to-end pipeline. Both regimes share the same verification backbone: facets define what must be supported, the verifier supplies the support signal, and selection decides what evidence is worth paying for under B_{ctx}. Figure 2 shows Safe-Cover's certified trace on a two-hop query.

3.1 Facets: Auditable Reasoning Requirements

A facet miner maps q to $F(q) = \{f_1, \ldots, f_m\}$, where each facet is intended to be checkable against a single passage. Facets are typed (e.g., ENTITY, RELATION, BRIDGE-HOP1, BRIDGE-HOP2, TEMPORAL, NUMERIC) to support stratified calibration and targeted diagnostics. Each facet includes (i) a hypothesis/claim template and (ii) a deterministic shortlisting key (anchors, triggers, or entity bindings) that defines exactly which passages will be tested by the verifier, ensuring the process remains auditable. Multi-hop queries may introduce placeholder facets whose hypotheses are not meaningful until a binding value is available. To avoid scoring non-instantiated hypotheses, placeholder facets are skipped during the initial scoring pass and evaluated only after instantiation (e.g., after Hop-1 binds an intermediate entity that populates a Hop-2 template). This implementation detail prevents spurious failures caused by testing ill-formed hypotheses. Facet mining is the functional bottleneck for certified behavior: the system can only certify what it can express as facets. Accordingly, all guarantee statements are conditional on the mined set $F(q)$ and the logged contract used to test it.

3.2 Candidates and Cost Model

A retriever produces a candidate pool $P = \{p_1, \ldots, p_n\}$. Each passage p has a nonnegative cost $c(p)$ representing the token cost of serializing that evidence into the prompt. We impose a hard evidence cap B_{ctx} and require $\sum_{p \in S} c(p) \leq B_{ctx}$ for any selected set S. Costs count evidence tokens only; generation tokens are tracked separately. The cost model and candidate pool form part of the immutable selection contract for the certified regime.

3.3 Verification as fixed tests and selection-conditional p-values

Deterministic shortlisting defines the tested set. Verifying every pair (p, f) is computationally wasteful and statistically unstable. For each facet f, a deterministic shortlister selects up to T_f passages from P using fixed rules and fixed tie-breaks. Only these shortlisted pairs are scored by the verifier. This turns "what was tested" into a declared, replayable object—a prerequisite for both valid multiple-testing control and external auditing.

Event of interest and verifier scores. For each shortlisted pair (p, f), a verifier produces a score $s(p, f) \in R$ intended to correlate with the event $\Sigma(p, f)$: "passage p is sufficient evidence for facet f". In practice, the verifier may be two-stage (lexical gate followed by an NLI score); the score $s(p, f)$ is the final verification statistic passed to calibration. This is the key semantic point: the p-value measures sufficiency for the facet, not merely whether a string appears.

Selection-conditional conformal p-values. Crucially, verifier score distributions are artifacts of the retrieval policy. We therefore calibrate under the same policy used at test time. Each shortlisted (p, f) is assigned to a bin $b = b(p, f)$ and mapped to a conformal p-value using a negative calibration pool $\mathcal{N}_b$ generated by replaying the identical

selection logic:

$$\pi(p, f) = \frac{1 + \sum_{u \in \mathcal{N}_b} I[s(u) \geq s(p, f)]}{|\mathcal{N}_b| + 1}, \quad (1)$$

with randomized tie-handling when scores are discrete. These p-values are valid only under the replayed contract.

Mondrian binning and feasibility. To reduce heterogeneity, we use Mondrian calibration with a bin key that includes at least facet type and a passage length bucket:

$$b(p, f) = \big(\text{canonical\_type}(f),\ \text{len\_bucket}(p)\big). \quad (2)$$

Deterministic conformal p-values have a floor $1/(|\mathcal{N}_b| + 1)$. If the target threshold falls below this floor, certification is mathematically impossible. We apply a fixed bin-merging rule to reach a minimum effective pool size, logging merge depth to the replay record. If feasibility still cannot be met, the certified regime does not guess—it returns an explicit reason for infeasibility.

3.4 Certified regime: Safe-Cover as risk-controlled min-cost facet cover

The certified regime converts calibrated facet tests into a fixed set system and solves a budgeted min-cost cover. It guarantees one of two outcomes: a certified evidence set with per-facet support certificates (and an answer extracted from certifying passages), where the probability of any false facet-support certificate is bounded by query, or a machine-checkable abstention. Figure 2 provides a concrete trace of this process.

Episode contract (frozen knobs). Within an episode (one query execution), all parameters defining the tested set and thresholds are frozen and logged: retriever snapshot and candidate cap; deterministic shortlisting rules and T_f; verifier version; binning scheme; calibration pools; evidence serialization rules; and the evidence budget B_{ctx}. No mid-episode adaptation is permitted. This is what makes the guarantee checkable: if any contract element changes, certificate validity is void.

Bonferroni allocation and fixed coverage sets. We allocate a query-level error budget α_{query} across facets and tests:

$$\alpha_f = \frac{\alpha_{\text{query}}}{|F(q)|}, \qquad \bar{\alpha}_f = \frac{\alpha_f}{T_f}. \quad (3)$$

A facet is covered by passage p when $\pi(p, f) \leq \bar{\alpha}_f$. This induces a fixed coverage set per passage:

$$C(p) = \{f \in F(q) \ :\ \pi(p, f) \leq \bar{\alpha}_f\}. \quad (4)$$

Because shortlisting and thresholds are frozen, $C(p)$ is fixed within the episode, restoring classical set-cover semantics.

Budgeted min-cost cover and reproducible greedy selection. Safe-Cover solves a min-cost cover under a hard evidence budget:

$$\begin{aligned} \min_{S \subseteq P} \quad & \sum_{p \in S} c(p) \\ \text{s.t.} \quad & \bigcup_{p \in S} C(p) \supseteq F(q), \\ & \sum_{p \in S} c(p) \leq B_{\text{ctx}}. \end{aligned} \quad (5)$$

We use a deterministic greedy algorithm that maximizes newly covered facets per unit cost, ensuring exact reproducibility.

Winning-passage alignment and answer extraction. Certification is useless if the answer is generated from an unverified context. For each facet, the system records a winning passage (typically the passage with the smallest p-value among candidates that passed the threshold). Answer extraction is aligned to these winning passages: typed extractors run on the certifying evidence, and the system does not silently substitute unrelated context if extraction fails. This resolves a standard failure mode where a certificate is computed on one piece of evidence, but the answer is generated from another.

Abstention certificates and reason codes. If any required facet has no covering passage, Safe-Cover returns a `NO_COVER` outcome. If coverage exists in principle but cannot be achieved under the remaining budget, Safe-Cover returns `INFEASIBLE_BUDGET` (optionally supported by a dual lower bound on remaining cover cost). If threshold feasibility fails due to bin floors after merging, Safe-Cover returns `INFEASIBLE_PVALUE`. All abstentions are normalized to a fixed output token (e.g., `ABSTAIN`) for evaluation integrity.

Certificate payload and query-level statement. A facet-support certificate includes the facet ID, winning passage ID, p-value, threshold, bin key, shortlist metadata, and version hashes. Under the frozen contract and Bonferroni allocation, the probability of emitting any false facet-support certificate within a query is bounded by α_{query}.

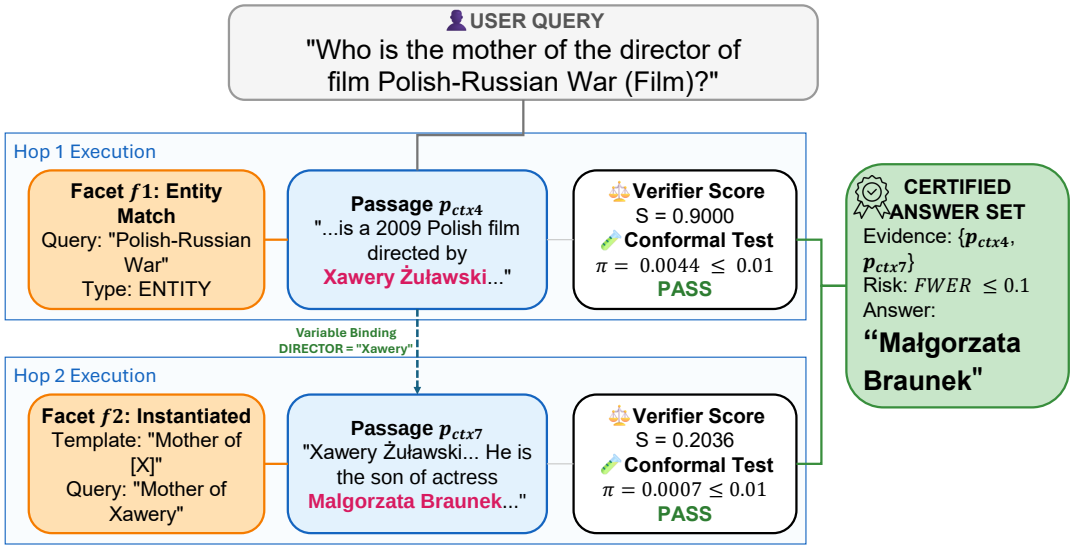

Figure 2: Safe-Cover trace on a two-hop query: Hop-1 binds an intermediate entity, Hop-2 instantiates a dependent facet, and passing tests yield certificates and a certified evidence set.

3.5 Fallback tier: Non-certified provenance-checked answering

When queries fall outside the certifiable regime—either because facet mining fails or Safe-Cover abstains—the system can optionally invoke a non-certified fallback tier that answers over the complete candidate pool. This tier enforces strict provenance checking but does not emit certificates or claim bounded error rates. Safe-Cover results reported in Section 4 include only certified outcomes.

3.6 Frontier regime: Pareto-Knapsack

The frontier regime drops per-query certification and instead optimizes a quality–cost objective under the same evidence cap. It treats verifier scores as continuous utility signals and greedily packs the context window to maximize evidence strength per token. Because this mode may employ adaptive mechanisms (e.g., query rewrites), it reports empirical quality–cost trade-offs without per-query validity claims. Figure 3 details the complete Pareto pipeline: facet mining and retrieval produce a candidate pool, batch scoring assigns verifier-derived utilities, and lazy greedy packing selects passages under the evidence budget B_{ctx} by maximizing marginal utility per cost. The system enforces strict provenance checking—answers must be extractable from the selected evidence—and abstains when provenance fails.

3.7 Operational safeguards and diagnostics

The system logs certificate payloads, shortlist sizes, bin depths, near-threshold counts, abstention codes, and version metadata sufficient for replay. These signals support debugging (coverage sparsity, verifier discrimination) and operational monitoring. If drift alarms trigger—indicating the test-time distribution has diverged from the calibration pool—the certified regime fails closed until re-calibration is performed.

4 Experiments

TRIDENT targets a regime where multi-hop questions require evidence, but evidence is expensive: both context length and latency must be actively controlled. We therefore evaluate TRIDENT as a quality–cost system rather than a pure-accuracy model. Across all experiments, we hold the generator and upstream retrieval/verifier stack constant; we vary only the serving regime (Pareto-Knapsack vs. Safe-Cover), the evidence budget, and (for Safe-Cover) the certification parameters. This isolates the impact of evidence selection from retrieval quality and generation capacity.

4.1 Experimental setup

Datasets. We evaluate on three multi-hop QA benchmarks: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). Each requires aggregating evidence across multiple passages—exactly the

Method / Config	Model	2Wiki				HotpotQA				MuSiQue			
		EM (%)	F1 (%)	EvTok	Lat (ms)	EM (%)	F1 (%)	EvTok	Lat (ms)	EM (%)	F1 (%)	EvTok	Lat (ms)
TRIDENT (Pareto-Knapsack; evidence-budget sweep)													
Pareto-400	Llama-3-8B-Instruct	25.95	33.83	251.39	3985.74	43.80	55.80	348.01	2893.70	15.02	26.93	344.75	1898.56
Pareto-500	Llama-3-8B-Instruct	28.16	36.25	331.48	5650.08	45.30	58.22	446.22	2541.01	19.21	30.74	446.30	2829.25
Pareto-1000	Llama-3-8B-Instruct	33.81	42.45	631.08	5476.45	50.23	63.68	874.63	2569.91	20.16	33.54	847.73	793.29
TRIDENT (Safe-Cover; certified mode)													
Safe-2000 (equal)	Llama-3-8B-Instruct	34.70	43.61	869.07	4711.2	49.24	62.55	1251.46	6080.91	3.82	8.74	1899.39	11921.22
Safe-4000 (loose)	Llama-3-8B-Instruct	34.52	43.42	871.20	6307.85	48.89	62.27	1251.33	3806.56	4.53	10.09	1898.32	11904.13

Table 1: TRIDENT main results. EvTok: average evidence tokens passed to the generator. Lat: per-query end-to-end latency average (ms) over all queries, including abstentions. Latency and token accounting follow App. §A–C. Abstention, confidence intervals and additional protocol details are in the appendix.

setting where top-k truncation is brittle under a hard cap. HotpotQA and 2Wiki provide annotated supporting facts for selection analysis under budget pressure. MuSiQue is a harder stress-test: longer reasoning chains and less redundant evidence expose when selection or certification becomes feasibility-limited.

Models and decoding. All budget-matched comparisons use the same model (Llama-3-8B-Instruct (AI@Meta, 2024) or Qwen3-8B (Team, 2025)) with identical decoding parameters (temperature 0.0, consistent stopping criteria, same `max_new_tokens`). We apply uniform answer normalization and extraction rules across systems (App. §A) and report both generators throughout but interpret within-generator comparisons most heavily, since absolute latency and output behavior can differ across model families under identical caps in our serving setup.

Retrieval and budget enforcement. All TRIDENT variants share the same retriever, re-ranker, and verifier; differences arise solely from selection logic. For baselines, we log retrieval conditions and—when dataset-provided contexts are used—the exact ordering rule applied before truncation. This matters on 2Wiki especially: passage ordering can determine whether a bridge passage survives a hard cap or gets truncated away. We enforce deterministic ordering and report it verbatim (App. §B), and every run logs the final evidence list passed to the generator. We distinguish evidence tokens (EvTok) from total input tokens (TotTok) and enforce a hard evidence-token cap by truncating after selection and before generation using a deterministic policy (App. §B). We report EvTok and latency alongside EM/F1 because cost is central to the objective—accuracy gains that require $3\times$ the tokens are real but less interesting for budget-constrained deployment.

Serving regimes. Pareto-Knapsack optimizes quality against cost under a hard evidence cap. We sweep budgets (400/500/1000 tokens) to trace the quality–cost frontier; main results use a relaxed acceptance threshold, with Section 4.4 showing that tightening α primarily raises abstention rather than improving accuracy proportionally.

Safe-Cover is the certified regime, reported at two configurations (Safe-2000 and Safe-4000) corresponding to different evidence caps and certification stringency. All episode-frozen knobs (shortlisting policy, tests-per-facet T_f, binning key, threshold allocation, tie-breaks, and serialization format) are locked before selection and logged as part of the audit record. This is what makes the certificate replayable: if any contract element changes, the validity claim does not transfer.

Latency is measured end-to-end (retrieve $\rightarrow$ rerank $\rightarrow$ verify $\rightarrow$ select $\rightarrow$ answer/abstain), including abstentions, under fixed batching and cache policies (App. §A). When comparing latency in prose, we specify whether we reference mean or percentile values and cite the corresponding table.

4.2 Diagnostics for certified behavior

Safe-Cover's core claim is a query-time certificate under a replayable contract. That claim is only credible if failures are auditable rather than hidden behind aggregate metrics. The system logs three classes of diagnostic signals: (1) per-facet coverage status with abstention reason codes (`NO_COVER`, `INFEASIBLE_BUDGET`, `INFEASIBLE_PVALUE`), (2) calibration feasibility indicators (bin sizes, merge depth, threshold floors), and (3) verifier discrimination metrics. Figure 2 shows a representative execution trace: tests run on fixed (p, f) pairs, passing

Method / Config	Model	2Wiki					HotpotQA					MuSiQue				
		EM	F1	EvTok	TotTok	Lat (ms)	EM	F1	EvTok	TotTok	Lat (ms)	EM	F1	EvTok	TotTok	Lat (ms)
Llama-3-8B-Instruct Comparisons																
TRIDENT Pareto-500	Llama-3-8B	28.16	36.25	**331**	**878**	5,650	45.30	58.22	**446**	**1,006**	**2,541**	19.21	30.74	446	**920**	**2,829**
VanillaRAG-500	Llama-3-8B	26.13	29.26	361	1,002	2,559	30.81	39.61	461	1,193	2,686	5.13	9.16	**445**	1,155	2,912
TRIDENT Pareto-1000	Llama-3-8B	**33.81**	**42.45**	631	1,179	5,476	**50.23**	**63.68**	875	1,438	2,570	**20.16**	**33.54**	848	1,310	**793**
VanillaRAG-1000	Llama-3-8B	27.05	30.28	441	1,163	**2,531**	31.57	40.55	544	1,360	2,673	5.83	9.62	565	1,396	2,834
HippoRAG2	Llama-3-8B	27.42	31.32	569	1,472	3,509	37.79	47.82	593	1,488	4,185	15.35	23.95	663	1,639	7,835
Qwen3-8B Comparisons																
TRIDENT Pareto-500	Qwen3-8B	13.15	17.41	**332**	1,332	25,118	28.11	36.13	**451**	1,468	25,728	3.64	6.40	446	**1,427**	21,814
VanillaRAG-500	Qwen3-8B	20.27	21.98	361	**1,241**	8,549	27.08	33.77	461	**1,465**	**8,677**	3.14	5.21	**445**	1,463	**9,857**
TRIDENT Pareto-1000	Qwen3-8B	**36.93**	**43.06**	632	1,661	26,273	**44.14**	**56.12**	873	1,913	29,963	7.29	10.94	847	1,847	42,463
VanillaRAG-1000	Qwen3-8B	21.32	23.23	441	1,410	8,793	27.87	34.70	544	1,640	8,718	3.14	5.35	565	1,716	9,980
HippoRAG2	Qwen3-8B	21.14	23.75	518	1,716	**4,783**	34.77	44.01	573	1,896	22,621	**14.77**	**20.47**	625	2,221	53,329
Reference only																
Self-RAG	Self-RAG	2.74	17.44	877	935	886	14.56	30.68	1,261	1,319	959	1.65	9.16	1,168	1,235	1,053

Table 2: Baseline comparison across varying evidence budgets (approx. 500 and 1000 tokens). EM/F1 are in %. Lat is average latency in ms. EvTok denotes evidence tokens passed to the generator; TotTok denotes total input tokens. The best and second-best results in each column are highlighted in bold and underlined, respectively.

facets record their winning passages, and answers are extracted from those winners. If extraction fails or certified coverage is infeasible, the system returns `ABSTAIN` with a machine-readable reason code that enables post-hoc diagnosis of whether failures stem from evidence scarcity, calibration power limits, verifier discrimination, or budget constraints. Table 5 decomposes abstention reasons across datasets.

Auditability artifacts. Safe-Cover emits a replayable certificate payload for each passed facet (facet id, winning passage id, p-value/threshold, bin key, version hashes) plus machine-checkable outcome codes under the episode-frozen contract. If drift alarms trigger, the certified regime fails closed until re-calibration.

4.3 Main results

Tables 1 and 2 show that TRIDENT's gains come from treating budgeted RAG as a selection problem with explicit intermediate requirements, not from retrieving more text or relying on longer contexts. Under strict caps, the dominant failure mode of top-k truncation is not that relevant evidence is absent, but that the critical bridge passage gets dropped or diluted within the packed context.

Selection quality under a hard cap. At matched evidence budgets, verification-aware selection yields large gains over naive concatenation. On HotpotQA at a 500-token cap, Pareto-500 improves EM/F1 from 30.81/39.61 (VanillaRAG-500) to 45.30/58.22—a +14.49/+18.61 absolute gain, 47% relative improvement on both metrics—while using slightly fewer evidence tokens (446 vs. 461). Using the mean end-to-end latency reported in Table 2, Pareto-500 is also modestly faster (2,541ms vs. 2,686ms), consistent with selecting less redundant context.

On MuSiQue, at essentially the same evidence budget (446 vs. 445 EvTok), Pareto-500 improves EM/F1 from 5.13/9.16 to 19.21/30.74 (3.7× EM; 3.4× F1). On 2WikiMultiHopQA, gains are smaller but consistent: 28.16/36.25 vs. 26.13/29.26 EM/F1 while using fewer evidence tokens (331 vs. 361). The pattern throughout: Pareto improves accuracy by packing better evidence, not by spending more tokens.

The budget knob behaves predictably. Sweeping the evidence budget yields a monotone, interpretable operating curve (Table 1): increasing the allowance from 400 to 500 tokens improves EM/F1, and increasing to 1000 improves further, with predictable cost increases and modest latency changes. This is exactly what a budgeted framework should provide—an operator can choose a target quality level and read off the corresponding token cost without redesigning the system at each operating point.

Abstention at Pareto-500 remains low but nonzero (HotpotQA 1.54%, 2Wiki 2.12%, MuSiQue 6.04%), consistent with fail-closed behavior: when evidence cannot support a reliable output under the cap, the system abstains rather than fabricating plausible citations.

Selection complements stronger retrieval. TRIDENT is not a retrieval substitute; it improves how a fixed candidate pool gets distilled under budget. Compared to HippoRAG (Table 2), Pareto-500 achieves higher accuracy while using fewer tokens and lower mean latency: on HotpotQA, +7.5 EM / +10.4 F1 with 25% fewer EvTok (446 vs. 593) and 39% lower latency (2,541ms vs. 4,185ms); on MuSiQue, +3.9 EM / +6.8 F1 with 33% fewer EvTok (446 vs. 663) and 64% lower latency (2,829ms vs. 7,835ms). Better retrieval and better selection compound rather than compete—expanding the candidate pool helps, but packing the right subset into a tight context window remains decisive.

Safe-Cover: trading cost for auditability. The certified Safe-Cover regime answers only when it can certify facet coverage under an episode-frozen contract, otherwise emitting ABSTAIN with an explicit reason code. Where certification is feasible, Safe-Cover remains accurate (HotpotQA 49.24/62.55 EM/F1; 2Wiki 34.70/43.61 in Table 1), but uses substantially more evidence tokens—the intended cost of certifying coverage rather than merely optimizing average accuracy. For applications where a wrong answer is worse than no answer, this trade-off makes sense.

The MuSiQue collapse. On MuSiQue, Safe-Cover achieves very low EM/F1 despite large evidence budgets. This does not invalidate the framework; it exposes where certification becomes power-limited. Safe-Cover can only certify what the verifier can separate and what calibration can support under bin floors and multiple-testing thresholds. Safe-Cover's failure mode is itself auditable—it fails closed with explicit reason codes. A system that confidently hallucinated answers would score higher on EM but be far less trustworthy in deployment.

4.4 Ablations and analysis

Our ablations test whether Pareto's gains rely on brittle implementation choices—a fragile reranker ordering, a knife-edge utility at $p = \alpha$, or a narrow threshold setting—or whether they reflect a stable interaction between verification signals and budgeted selection. The distinction matters: brittle gains evaporate when conditions shift; stable gains transfer.

Threshold sensitivity. Sweeping the relaxed acceptance threshold α (detailed plots in Appendix A) shows that tightening α sharply increases abstention—the system cannot assemble a budget-feasible cover—while EM/F1 changes more modestly. The evidence budget, not the threshold, is the binding constraint on token consumption. We use a relaxed setting in the main results to prioritize answering when evidence is plausibly sufficient; operators with stricter requirements can tighten α and accept higher abstention.

Config		EM / F1	Abstain %	EvTok	TotTok	Lat p50
α=0.01 (Strict)	HotpotQA	43.67% / 56.37%	10.93%	454.25	1017.32	514.95
α=0.02	HotpotQA	44.39% / 57.57%	8.17%	454.40	1018.23	506.62
α=0.05	HotpotQA	44.63% / 56.53%	16.21%	454.20	1015.11	516.44
α=0.1	HotpotQA	44.77% / 57.37%	6.89%	454.05	1012.57	512.94
α=0.2	HotpotQA	44.94% / 57.39%	4.17%	452.65	1016.12	1030.03
α=0.5	HotpotQA	45.56% / 58.45%	1.54%	451.02	1011.33	1166.79
α=0.6 (Loose)	HotpotQA	45.30% / 58.22%	1.54%	446.22	1006.11	467.80
No-Rerank (α=0.6)	HotpotQA	45.94% / 58.08%	1.54%	450.51	1011.62	2275.23
Soft Sigmoid (α=0.6)	HotpotQA	46.14% / 59.20%	1.54%	466.90	1025.37	531.21
Soft Sigmoid (α=0.6)	2Wiki	28.90% / 36.81%	2.12%	334.58	879.78	1148.53
Soft Sigmoid (α=0.6)	Musique	17.28% / 28.62%	6.04%	460.32	934.35	1504.08

Table 3: Pareto ablations under a fixed evidence budget. Different α configurations sweeps the relaxed threshold α (HotpotQA), config No-Rerank removes reranking, and config soft sigmoid replaces binary marginal gain with a soft-sigmoid utility. We report EM/F1, abstention, EvTok, and Lat_{50}.

Reranking and utility design. Removing reranking leaves HotpotQA EM/F1 and abstention essentially unchanged (Table 3), while increasing latency—Pareto's 500-token gains are not artifacts of fragile reranker-induced ordering, as reranking primarily improves efficiency by reducing noise in the candidate list. This suggests the selection logic is robust to upstream perturbations. Similarly, replacing the binary utility with a soft-sigmoid score (utility_tau of 0.05) yields comparable and slightly improved HotpotQA EM/F1 (46.14/59.20 vs. 45.30/58.22) at the same abstention rate (1.54%), with a modest EvTok increase. Pareto does not depend on a knife-edge decision boundary at $p = \alpha$; smoothing near-threshold support can help the greedy solver in borderline cases without changing the semantic definition of coverage.

5 Conclusion

We formalize budgeted RAG as typed facet verification under hard token limits. Pareto-Knapsack traces a predictable accuracy–cost curve, improving HotpotQA EM by 47% relative at matched budgets through principled passage packing rather than expanded retrieval. Safe-Cover offers a certified alternative: verifiable support certificates under frozen contracts, or auditable abstention when certification fails.

Limitations

Contract-scoped validity. Safe-Cover certificates are claimed only under the logged selection contract: the same retriever snapshot, deterministic shortlisting policy (including tie-breaks and T_f), verifier, calibration bins, and evidence serialization rules used during calibration must be replayed at test time. If any element changes, certificate validity is not claimed.

Sensitivity to verifier and facet design. Certification depends on the verifier's ability to separate sufficient from insufficient support at the facet granularity. Poorly specified facets, weak verifier discrimination, or sparse calibration bins can make certification infeasible even when a correct answer exists in the candidate pool.

Conservatism under diffuse evidence. Safe-Cover can be conservative on tasks where support is distributed across passages or does not align cleanly with single-passage facet checks, as seen on MuSiQue. This is less a failure of the abstraction than a reminder that auditable certification fundamentally trades coverage for strictness.

Scope of evaluation. We focus on multi-hop QA benchmarks and report accuracy, evidence tokens, and latency. Broader tasks (summarization, long-form synthesis) may require different facet designs and verification signals.

References

AI@Meta. 2024. Llama 3 model card.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint*. ArXiv:2107.07511 [cs].

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint*. ArXiv:2310.11511 [cs].

Jiale Deng, Yanyan Shen, Ziyuan Pei, Youmin Chen, and Linpeng Huang. 2025. Influence Guided Context Selection for Effective Retrieval-Augmented Generation. *arXiv preprint*. ArXiv:2509.21359 [cs].

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. *arXiv preprint*. ArXiv:2405.14831 [cs] TLDR: HippoRAG is introduced, a novel retrieval framework inspired by the hippocampal indexing theory of human long-term memory to enable deeper and more efficient knowledge integration over new experiences and can tackle new types of scenarios that are out of reach of existing methods.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv preprint*. ArXiv:2002.08909 [cs].

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

Gautier Izacard and Edouard Grave. 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv preprint*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint*.

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. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How Hard is Automatic Attribution Evaluation? *arXiv preprint*. ArXiv:2402.15089 [cs].

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. 2025. TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6599–6612, Suzhou, China. Association for Computational Linguistics.

Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2025. FIT-RAG: Black-Box RAG with Factual Information and Token Reduction. *ACM Transactions on Information Systems*, 43(2):1–27. TLDR: A novel black-box RAG framework which utilizes the factual information in the retrieval and reduces the number of tokens for augmentation, dubbed FIT-RAG, which achieves both superior effectiveness and efficiency.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv preprint*. ArXiv:2305.14251 [cs].

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Jingjin Wang and Jiawei Han. 2025. PropRAG: Guiding Retrieval with Beam Search over Proposition Paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6223–6238, Suzhou, China. Association for Computational Linguistics.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025. InfoGain-RAG: Boosting Retrieval-Augmented Generation through Document Information Gain-based Reranking and Filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7201–7215, Suzhou, China. Association for Computational Linguistics.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. *arXiv preprint*. ArXiv:2009.12756 [cs].

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. *arXiv preprint*. ArXiv:2401.15884 [cs].

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

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé. 2021. Multi-Step Reasoning Over Unstructured Text with Beam Dense Retrieval. *arXiv preprint*. ArXiv:2104.05883 [cs].

A Appendix: Deferred System Details

A. Evaluation protocol, statistical treatment, and latency definition

Confidence intervals. Table 4 provides 95% confidence intervals for all systems and datasets reported in Tables 1 and 2. These intervals quantify statistical uncertainty and support the significance claims in Section 4.3.

Statistical significance. Non-overlapping confidence intervals indicate statistically significant differences at $\alpha = 0.06$. All TRIDENT Pareto comparisons against VanillaRAG baselines show non-overlapping intervals for both EM and F1, confirming that the gains reported in Section 4.3 are statistically robust.

Method	Dataset	EM 95% CI	F1 95% CI
TRIDENT Pareto-500 (Llama-3-8B)			
Pareto-500	HotpotQA	[44.2, 46.4]	[57.1, 59.4]
Pareto-500	2Wiki	[27.3, 29.0]	[35.3, 37.2]
Pareto-500	MuSiQue	[17.5, 21.0]	[28.8, 32.7]
TRIDENT Pareto-1000 (Llama-3-8B)			
Pareto-1000	HotpotQA	[49.1, 51.3]	[62.5, 64.8]
Pareto-1000	2Wiki	[32.9, 34.7]	[41.4, 43.5]
Pareto-1000	MuSiQue	[18.3, 22.0]	[31.5, 35.6]
TRIDENT Safe-Cover (Llama-3-8B)			
Safe-2000	HotpotQA	[48.1, 50.4]	[61.3, 63.8]
Safe-2000	2Wiki	[33.9, 35.5]	[42.5, 44.7]
Safe-2000	MuSiQue	[2.9, 4.8]	[7.2, 10.3]
Safe-4000	HotpotQA	[47.7, 50.0]	[61.0, 63.4]
Safe-4000	2Wiki	[33.7, 35.4]	[42.3, 44.5]
Safe-4000	MuSiQue	[3.5, 5.6]	[8.5, 11.7]
Baselines (Llama-3-8B)			
VanillaRAG-500	HotpotQA	[29.8, 31.8]	[38.4, 40.8]
VanillaRAG-500	2Wiki	[25.3, 26.9]	[28.2, 30.3]
VanillaRAG-500	MuSiQue	[4.1, 6.2]	[7.8, 10.5]
VanillaRAG-1000	HotpotQA	[30.5, 32.6]	[39.3, 41.8]
VanillaRAG-1000	2Wiki	[26.2, 27.9]	[29.3, 31.3]
VanillaRAG-1000	MuSiQue	[4.8, 6.9]	[8.3, 11.0]
HippoRAG	HotpotQA	[36.7, 38.9]	[46.6, 49.0]
HippoRAG	2Wiki	[26.6, 28.3]	[30.3, 32.4]
HippoRAG	MuSiQue	[13.7, 17.0]	[22.0, 25.9]

Table 4: 95% confidence intervals for EM and F1 metrics. Wilson score intervals for EM; stratified bootstrap (B=1000) for F1. Non-overlapping intervals indicate statistically significant differences at $\alpha = 0.05$.

Latency definition and inclusion policy. Latency is measured end-to-end per query over the full pipeline: retrieve $\rightarrow$ rerank $\rightarrow$ verify/score $\rightarrow$ select $\rightarrow$ generate (or abstain). All reported latency percentiles include abstentions. All methods use the same batching configuration for retrieval, reranking, and verification. We keep the caching policy fixed within a run and compute Lat_{50}/Lat_{90}/Lat_{95} under that policy.

Token accounting. We distinguish (i) evidence tokens (EvTok), the number of tokens contributed by the evidence passages passed to the generator, and (ii) total input tokens (TotTok), the full generator input including the fixed prompt template and evidence. Both are computed using the generator tokenizer. Completion lengths are standardized via a shared decoding configuration across all methods.

B. Baseline fairness and budget enforcement

Evidence-token budget. A budget of B evidence tokens denotes a hard cap on the concatenated evidence text passed to the generator after a method's final selection step. Any method that constructs an explicit evidence context is forced to respect this cap.

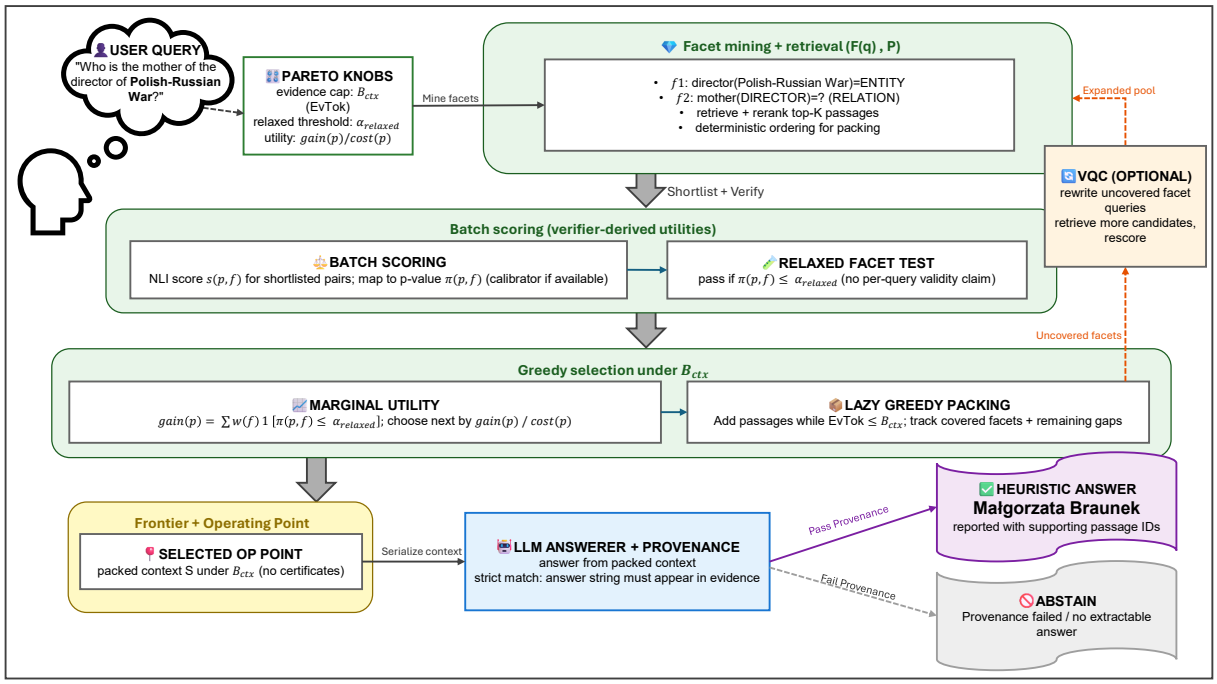

Figure 3: Pareto-Knapsack selection pipeline. Facets and candidates flow through batch scoring with relaxed thresholds ($\alpha_{relaxed}$), greedy selection maximizes marginal utility per cost under the evidence budget B_{ctx}, and the LLM answerer enforces strict provenance checking. Optional VQC (Variable Query Completion) can retrieve additional candidates for uncovered facets.

Deterministic truncation policy. If the selected evidence exceeds B, we truncate deterministically: passages are concatenated in the method's final priority order until the cap is reached. The final passage is truncated at the exact token boundary. This enforces budget compliance without altering the method's internal ranking logic.

Generator and candidate parity. Unless otherwise stated, head-to-head comparisons use the same generator model and decoding parameters (temperature, top-p, `max_new_tokens`) to isolate the impact of selection from generation capacity. When a baseline uses the shared retrieval stack, it receives the same candidate pool and reranker ordering as the frontier regime. When a baseline (e.g., HippoRAG) uses its own retrieval strategy, we treat it as a retrieval-strength reference but still enforce the same evidence-token cap. All runs log the final evidence list actually passed to the generator.

C. Abstention definition and reporting

Definition. An abstention occurs when the system intentionally returns a dedicated token because it cannot produce an evidence-conditioned answer under the configured constraints. Abstention rates are reported as a percentage of evaluated queries. In the certified regime, abstentions carry specific reason codes: `NO_COVER` (some required facet has no passing passage), `INFEASIBLE_BUDGET` (a valid cover exists but exceeds the cap), or `INFEASIBLE_PVALUE` (calibration bins too sparse to support the required threshold).

Dataset	Answered (%)	No Cover (%)	Infeasible Budget (%)	Other (%)
HotpotQA	98.5	1.1	0.15	0.25
2Wiki	97.9	1.3	0.3	0.5
MuSiQue	94.0	2.7	0.8	2.5

Table 5: Abstention reason distribution for Safe-Cover 2000 regime. No Cover: some required facet has no passing passage. Infeasible Budget: dual_lower_bound > evidence_token_cap. Other: Infeasible p-value, system error, etc.. MuSiQue's high No Cover rate reflects the difficulty of certifying evidence for longer reasoning chains.

Counts at the Pareto-500 α = 0.6 operating point. Table 6 details the raw abstention counts for the primary operating point reported in the main paper.

D. Episode contract and replay requirements

Contract elements. Certificate validity is claimed only under a replayed selection contract, as described in Section 3.4. The contract includes,

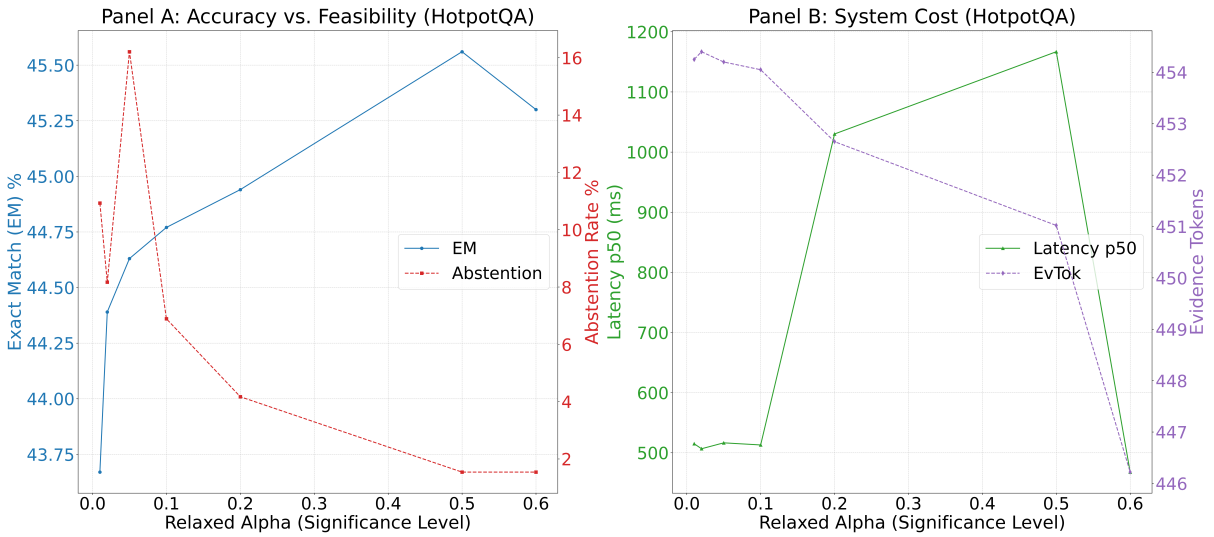

Figure 4: Pareto threshold sensitivity on HotpotQ.

Dataset	Method	Abstained / N	Rate (%)
HotpotQA	Pareto-500	114 / 7405	1.54
2WikiMultiHopQA	Pareto-500	266 / 12576	2.12
MuSiQue	Pareto-500	146 / 2417	6.04

Table 6: Abstention rate summary at the Pareto-500 operating point.

at minimum: retriever snapshot and candidate cap; deterministic shortlisting rules and tie-breaks (including T_f); verifier version and scoring configuration; binning key and bin-merging policy; calibration pools (or their hashes); evidence serialization format and tokenization; and the evidence budget B_{ctx}. All contract elements are logged as version hashes plus runtime knobs.

Failure behavior. If any contract hash mismatches (e.g., index snapshot changes, verifier weights change), certificates are declared invalid and the certified regime is disabled. The system fails closed—falling back to abstention or non-certified operation—rather than emitting certificates under unknown conditions. This keeps the guarantee honest.

E. Shift monitoring and conservative fallback

Drift signals. We log per-bin score summaries and near-threshold counts to detect distributional shift in verifier scores. This monitoring is treated as an operational safeguard rather than a formal proof obligation within the core guarantee.

Fallback policy. If drift alarms trigger, the certified regime suspends certificate emission until re-calibration is performed under the updated contract. During this window, the system may continue to answer in the non-certified regime but does not claim the certified error bound.

F. Deferred components

Verifier-driven query rewriting. As shown in Figure 3, typed rewrites can improve recall by retrieving additional candidates for uncovered facets. However, because rewriting changes the candidate distribution—and thus invalidates calibration conditions—we restrict rewriting to non-certified operation in the main protocol. In Pareto mode, VQC operates as an optional expansion layer that rewrites uncovered facet queries, retrieves more candidates, and rescores them before final selection.

Long-run budget control. A lightweight controller can manage token budgets across a stream of queries (e.g., amortization across sessions), but this is orthogonal to the per-query certification focus of this paper.

G. Label noise sensitivity

If a fraction ϵ of calibration negatives are mislabeled, conformal p-values become more conservative than intended. We optionally include a sensitivity analysis that inflates denominators to model this noise, reporting trends across $\epsilon \in \{0, 0.01, 0.05, 0.10\}$. This analysis aids deployment planning but is not required for the core validity claims.

H. Relaxed-α sweep: feasibility vs. accuracy

Figure 4 analyzes the trade-off between strictness and feasibility on HotpotQA by varying the relaxed

acceptance threshold α.

Panel A shows that the dominant effect of tightening α is on feasibility: abstention rates rise sharply as the system struggles to assemble a budget-feasible cover, while EM changes more gradually.

Panel B confirms that evidence usage (EvTok) remains essentially flat across values of α. The changes in performance are not driven by "spending" more tokens—the evidence cap is binding and stable. Latency variations instead reflect pipeline dynamics, specifically how easily the greedy solver finds passing support under stricter thresholds.