TRIDENT: Post-Selection Evidence Accountability for Token-Efficient RAG

Wutong Zhang[1] **Richard Liu**[1] **Sizhe Zhang**[1]
Hefeng Zhou[1] Kaixiang Wang[1] Celimuge Wu[2]
Wei Zhao[3] Jiong Lou[1,4*] Jie Li[1,4]
[1]Shanghai Jiao Tong University, China
[2]The University of Electro-Communications, Japan
[3]Shenzhen University of Advanced Technology, China
[4]State Key Laboratory of Submarine Geoscience, China

Abstract

RAG systems often verify citations only after retrieval and shortlisting have selected which passages will be tested. A verifier calibrated on random passage–claim pairs is therefore applied to a different population: shortlisted pairs chosen by the retrieval policy. This makes support verification a post-selection inference problem. We introduce TRIDENT, a contract-scoped framework for post-selection evidence accountability. Safe-Cover emits structural-facet support certificates with query-level Bonferroni FWER control, or abstains when the logged contract cannot certify support. Pareto-Knapsack relaxes certification and uses the same frozen facet table as a token-efficient multi-hop QA selector. Across HotpotQA, 2Wiki, and MuSiQue, TRIDENT controls false certificates under deployed contracts, identifies higher-quality answer-linked slices when certificates are emitted, and improves the matched-budget multi-hop QA frontier. The accountable unit is therefore not a retrieved passage or generated citation, but a support claim calibrated under the policy that selected it.

1 Introduction

Retrieval-augmented generation [Lewis et al., 2020, Karpukhin et al., 2020, Izacard and Grave, 2021] verifies citations only after retrieval and shortlisting have selected which passages will be tested. The ordering is not innocent: a verifier calibrated on random passage–claim pairs need not stay valid once a retriever preferentially retains high-scoring or lexically aligned candidates—the resulting citation is a post-selection inference [Berk et al., 2013, Lee et al., 2016, Fithian et al., 2017, Tibshirani et al., 2016]. Empirically, replaying our deterministic top-T shortlister on held-out nulls and applying a pre-selection calibrator inflates Type-I rejection to 1.7–$3.6\times$ the nominal level at $\tau{=}0.05$ on HotpotQA, 2Wiki, and MuSiQue (§4.2). The rejected nulls are precisely those that look most plausible to an auditor (Figure 1); a stronger verifier does not fix this, because the invalid population shift is induced upstream by the shortlister.

The post-selection lens has a second consequence beyond the Type-I inflation above. Multi-hop questions require evidence for several typed structural facets (e.g., a discoverer, an institution, a location), but a budgeted top-k selector can spend the evidence budget on passages that repeat an easy facet while dropping the bridge facet that justifies the answer. This is a facet-coverage failure of selection, parallel to the statistical calibration failure above. A support system must therefore both (i) cover the mined structural facets under budget and (ii) test each claimed facet support against a

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

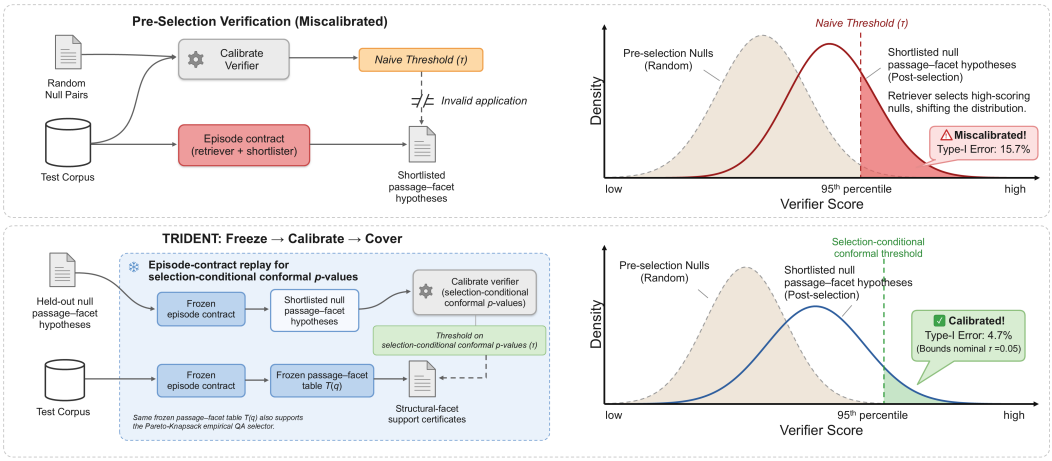

Figure 1: **Why RAG support verification is post-selection inference.** Retrieval shortlisting changes the null distribution of tested passage–facet hypotheses, making pre-selection verifier thresholds anti-conservative. TRIDENT replays the same shortlister during calibration, so support tests are calibrated on the selected null distribution. Per-dataset Type-I numbers in §4.2.

null distribution generated by the same selection policy; TRIDENT addresses both by making the selected passage–facet hypotheses explicit under a frozen episode contract.

Problem statement and system model. Given a query q, a candidate pool P produced by a frozen retriever, a typed structural facet set $F_{\text{struct}}(q)$, a passage cost $c(p)$, and a hard evidence-token budget B_{ctx}, we seek a selection function $\mathcal{S} : q \mapsto S \subseteq P$ with $\sum_{p \in S} c(p) \leq B_{\text{ctx}}$ together with a per-facet support test that controls a query-level false-certificate event under the policy that produced S. The system must (i) induce a tested-hypothesis family by deterministic shortlisting; (ii) calibrate p-values for that selected family rather than for the unconditional null; and (iii) compose the calibrated tests into a query-level coverage decision under B_{ctx}. We package the components that determine $\mathcal{S}$ and the per-facet test into an *episode contract* $E =$ (retriever, miner, shortlister, verifier, calibration pool, B_{ctx}, tie-break seed); freezing E makes the tested family replayable and the support tests selection-conditional. Notation, the formal p-value, and the FWER bound are in §3.

TRIDENT: *Freeze* → *Calibrate* → *Cover*. TRIDENT operationalizes the contract through two complementary modes that share one frozen passage–facet table $T(q)$. *Freeze* logs E. *Calibrate* replays the same shortlister on held-out nulls to obtain selection-conditional conformal p-values $\pi(p, f)$ per typed facet. *Cover* converts those calibrated tests into query-level decisions:

- **Safe-Cover (validity regime).** Solves a budgeted min-cost cover and emits structural-facet support certificates with query-level Bonferroni FWER control, or abstains with a typed reason code.
- **Pareto-Knapsack (efficiency regime).** Drops certification and packs evidence by marginal facet gain per token under B_{ctx}, achieving the highest mean matched-budget multi-hop F1 among compression, reranking, and selector baselines, with fewer evidence tokens than the next-best baseline (§4.5). It does not inherit Safe-Cover's FWER guarantee.

Safe-Cover and Pareto-Knapsack share the same selected passage–facet table but answer different questions: Safe-Cover asks when selected evidence can be certified; Pareto-Knapsack asks how to spend a token budget when certification is disabled.

Contributions.

- **Post-selection evidence accountability for RAG.** We formalize RAG support verification as selected-hypothesis testing: retrieval and shortlisting determine which passage–facet

hypotheses are tested, so pre-selection verifier thresholds become anti-conservative on the tested distribution (Proposition 1; §4.2).

- **A frozen passage–facet table as shared representation.** TRIDENT logs an episode contract and constructs $T(q)$, a replayable table of selected passage–facet hypotheses with facet types, costs, verifier scores, and selection-conditional support statistics.
- **Safe-Cover: the validity regime.** Safe-Cover composes replay-calibrated conformal p-values (Theorem 1) with budgeted structural-facet cover to control query-level FWER over false structural support certificates (Theorem 2), or abstain when certification is infeasible. Audited false-certificate point estimates stay below the nominal $\alpha_{\text{query}}{=}0.10$ under rebuilt LLM-miner conformal pools.
- **Pareto-Knapsack: the efficiency regime.** Pareto-Knapsack addresses the facet-coverage failure of whole-query top-k selection by packing evidence according to marginal facet gain per token, using the same frozen table $T(q)$ as a utility table. At a 500-token cap it obtains mean multi-hop F1 41.74 versus RECOMP's 40.98, using 408 versus 487.52 mean evidence tokens; this mode targets token-efficient QA.

2 Related work

Prior work improves retrieval, compression, reranking, or attribution. TRIDENT asks a different question: *after retrieval and shortlisting have selected evidence, are the resulting support claims calibrated for that selected distribution?*

Selection-side methods: budget compression, planning, and multi-hop expansion. Token and context compressors [Jiang et al., 2023, Xu et al., 2024, Wang et al., 2023], rerankers [Xiao et al., 2024], and selector-based methods reduce evidence to fit a budget; iterative and planning-based retrievers [Trivedi et al., 2023, Jin et al., 2025, Asai et al., 2024, Guu et al., 2020, Yan et al., 2024], learned multi-step retrievers [Sorokin et al., 2026, Zhang et al., 2024], and graph-/proposition-indexed retrievers [Gutiérrez et al., 2024, Wang and Han, 2025, Xiong et al., 2021, Zhao et al., 2021, Li et al., 2024a] expand recall on bridge passages. All target answer accuracy, retrieval recall, or token utility rather than post-selection support validity, and each induces a selected hypothesis family without contract-matched calibration. They are therefore *upstream policies to which TRIDENT's replay calibration applies*, not substitutes for post-selection support testing; comparisons are not matched on evidence budget (Appendix B.1).

Attribution, faithfulness, and post-selection inference. Attribution [Min et al., 2023, Li et al., 2024b] and influence-based selection [Deng et al., 2025, Wang et al., 2025, Mao et al., 2025, Lu et al., 2025] ask whether generated claims are supported, often after generation; conformal prediction [Angelopoulos and Bates, 2022] provides marginal validity under exchangeability, and the selective-inference literature [Berk et al., 2013, Lee et al., 2016, Fithian et al., 2017, Tibshirani et al., 2016] formalizes the bias retrieval shortlisting induces on any pre-selection calibrator. TRIDENT defines the correct exchangeability object for RAG support testing—the shortlisted passage–facet pair under the frozen episode contract—and gives a selection-conditional, query-level error statement for the evidence that systems already emit.

3 Framework

TRIDENT has three primitives. Freeze defines the shortlisted passage–facet hypotheses by logging the retrieval and shortlisting policy. Calibrate replays that policy on held-out nulls so support p-values are valid for selected, not random, evidence. Cover converts those calibrated pairwise tests into query-level evidence decisions under a token budget.

Setup and notation: query q; candidate pool P; typed facets $F(q)$ partitioned into structural $F_{\text{struct}}(q)$ and optional facets; passage–facet pair (p, f); verifier score $s(p, f)$; deterministic shortlister $\mathrm{Shortlist}(f)$ keeping the top T_f pairs; passage cost $c(p)$; evidence budget B_{ctx}. The tested hypotheses are the shortlisted pairs, not all (p, f). We reserve $\pi(p, f)$ for the conformal p-value of the pair (p, f) (Eq. 1) and $\rho_{\text{cert}}^{\text{struct}}$ for the query-level certificate-coverage rate; the two are not

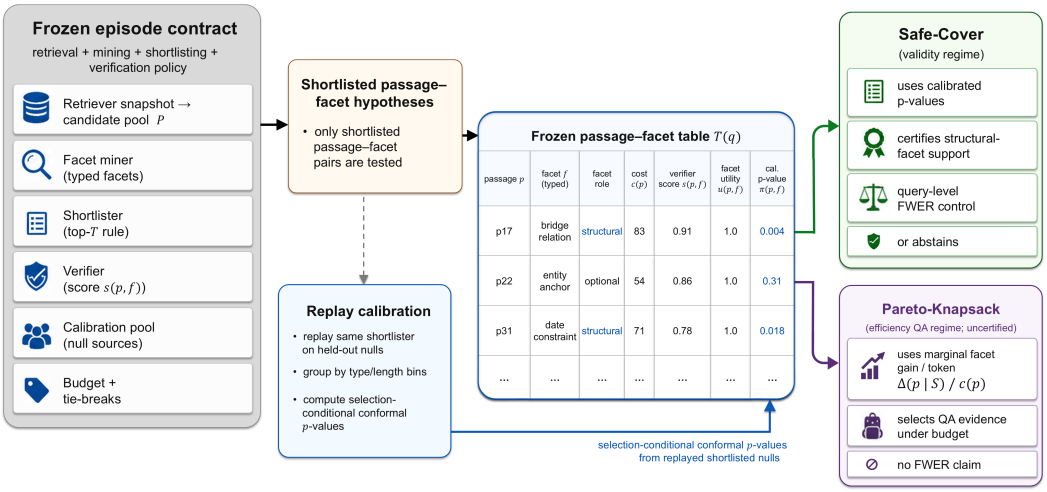

Figure 2: **TRIDENT's shared passage–facet representation.** A frozen episode contract induces the shortlisted passage–facet table $T(q)$. Safe-Cover reads $T(q)$ as a calibrated testing table and emits FWER-controlled structural support certificates or abstains. Pareto-Knapsack reads the same $T(q)$ as a utility table and selects evidence by marginal facet gain per token for efficient QA. The two regimes share representation but make different claims.

interchangeable. Operational details (placeholder hypothesis handling, candidate cost model, contract schema, diagnostic telemetry) are in Appendix C.

The central object in TRIDENT is a row of the frozen passage–facet table $T(q)$: a passage–facet hypothesis selected by the episode contract. Replay calibration assigns each row a selection-conditional conformal p-value $\pi(p, f)$ against a held-out null pool generated by the same shortlister, so calibration and coverage operate on the selected hypothesis distribution rather than on an unconditional pool of passage–claim pairs. Safe-Cover reads this row statistically, thresholding $\pi(p, f) \leq \bar{\alpha}_f$ to emit query-level FWER-controlled structural support certificates or abstain. Pareto-Knapsack reads the same row operationally, using its facet type, cost, and support signal to select high-utility evidence under a token budget. Figure 2 summarizes this shared representation: the two regimes differ in claim type, not in representation.

3.1 Freeze: episode contracts and the structural / optional split

A facet miner maps q to $F(q)$; each facet is typed (ENTITY, RELATION, BRIDGE-HOP, NUMERIC, TEMPORAL) and carries a hypothesis template and a shortlisting key. We split the typed space into *structural* $\mathcal{T}_{\text{struct}} = \{\text{BRIDGE}, \text{BRIDGE-HOP}_*, \text{RELATION}, \text{NUMERIC}, \text{TEMPORAL}\}$ and *optional* (typically ENTITY); only $F_{\text{struct}}(q)$ enters the Bonferroni denominator (optional facets get $\bar{\alpha}_f = 1$). All guarantees are conditional on $F_{\text{struct}}(q)$ and the logged contract.

Miner. In the main system, facets are produced by a contract-logged Llama-3-8B-Instruct facet planner whose prompt, model version, decoding seed, and JSON-normalization rules are frozen in the episode contract. We also evaluate a deterministic regex miner as a replay stress test and ablation. Theorem 2 applies conditional on the logged facet set and contract-matched calibration; swapping miners requires rebuilding the affected null pools.

An *episode contract* E freezes the components above plus the Mondrian binning key, evidence serializer, structural-type set, abstention/handoff flags, and tie-break seed; the full schema and a bit-identical replay statement are in Appendix C.

3.2 Calibrate: selection-conditional conformal facet tests

The hypotheses tested by TRIDENT are not all possible passage–facet pairs; they are the pairs that survive the frozen shortlisting rule. For each facet f the shortlister keeps up to T_f candidates, and only these are scored by a verifier $s(p, f) \in \mathbb{R}$ (lexical gate followed by an NLI score from a DeBERTa-v3 entailment model [He et al., 2023, Williams et al., 2018]). Shortlisting prefers high-scoring candidates, so the null score distribution of $s(p, f)$ conditional on $\{(p, f) \in \mathrm{Shortlist}(f)\}$ stochastically dominates the unconditional null:

Proposition 1 (Anti-conservativeness of pre-selection calibration)**.** *Fix bin b and a null pair (p, f). Let F_0 and F_0^{sel} denote the unconditional and selection-conditional null cdfs of $s(p, f)$, and* assume *$F_0^{\mathrm{sel}} \le F_0$ pointwise—i.e., the selection-conditional null first-order stochastically dominates the unconditional null. (Sufficient: a score-monotone top-T_f rule; necessary in general only that selection raises the upper tail of $s(p, f)$.) The pre-selection p-value $\tilde{\pi}(p, f) = 1 - F_0(s(p, f))$ satisfies $\Pr_{F_0^{\mathrm{sel}}}[\tilde{\pi}(p, f) \le \tau] \ge \tau$ for all $\tau \in [0, 1]$, with strict inequality when $F_0^{\mathrm{sel}} \ne F_0$ at the relevant threshold. For discrete verifier scores, the same conclusion holds using left limits and generalized quantiles in place of F_0^{-1}; strict inflation occurs at thresholds where the selection-conditional null places additional mass above the unconditional quantile. Proof in Appendix R; we exhibit the inflation empirically in Section 4.2.*

Toy intuition. If the shortlister keeps the maximum-scoring null among ten candidates, the selected score follows the maximum-order statistic of ten i.i.d. draws, not the unconditional null. A threshold calibrated to reject 5% of unconditional nulls will reject substantially more than 5% of selected nulls.

We therefore calibrate under the same policy used at test time. Each shortlisted (p, f) is assigned to a Mondrian bin $b(p, f) = (\text{canonical_type}(f), \text{len_bucket}(p))$ and mapped to a conformal p-value against a held-out negative pool $\mathcal{N}_b$ generated by replaying the identical shortlister. In our QA benchmarks, a null passage–facet pair is a shortlisted pair whose passage is not annotated as supporting the facet under the benchmark evidence labels. Benchmark annotations may miss alternative valid support. This benchmark-negative construction does not itself establish human semantic non-support; validity still requires exchangeability of the contract-defined null scores. The conformal p-value is

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

with randomized tie-breaks. Equation (1) uses the conservative upper-tail rank with ties counted against the test point. With randomized tie-breaking this equals the standard conformal rank p-value in expectation; without randomization it is conservative, which is sufficient for super-uniformity. Because $\mathcal{N}_b$ is generated under the same shortlister, exchangeability of $s(p, f)$ with $\{s(u) : u \in \mathcal{N}_b\}$ holds *conditional on* being shortlisted.

Theorem 1 (Marginal validity under Mondrian binning)**.** *If $s(p, f)$ is exchangeable with $\{s(u) : u \in \mathcal{N}_b\}$ conditional on b and the null, then for every $\tau \in [0, 1]$, $\Pr[\pi(p, f) \le \tau] \le \tau$. Proof in Appendix R.*

Bins whose held-out null pool falls below a fixed floor $N_{\min}$ are merged with their nearest length-bucket neighbour to keep the conformal p-value grid resolvable; merge depth is logged, and a bin that still cannot resolve $\bar{\alpha}_f$ after merging triggers `INFEASIBLE_PVALUE` rather than emitting a certificate. The full bin-merging rule, the heuristic-mode floor for runs without a Mondrian calibrator, and a sensitivity sweep over $(L, N_{\min})$ are in Appendices C and O.

3.3 Cover: risk-controlled structural facet cover

We allocate a query-level error budget α_{query} across the structural facet set and the per-facet test budget. If $F_{\text{struct}}(q) = \emptyset$, Safe-Cover emits a typed NO_STRUCTURAL_FACETS abstention and Eq. (2) is not evaluated; otherwise

$$\alpha_f = \begin{cases} \dfrac{\alpha_{\text{query}}}{|F_{\text{struct}}(q)|} & f \in F_{\text{struct}}(q), \\ 1 & f \notin F_{\text{struct}}(q), \end{cases} \qquad \bar{\alpha}_f = \begin{cases} \alpha_f / T_f & f \in F_{\text{struct}}(q), \\ 1 & f \notin F_{\text{struct}}(q). \end{cases} \tag{2}$$

Optional facets may contribute evidence utility and diagnostic certificates, but they are excluded from the FWER event and therefore do not consume the query-level error budget. A facet is covered by p when $\pi(p, f) \le \bar{\alpha}_f$, inducing a fixed coverage set $C(p) = \{f \in F(q) : \pi(p, f) \le \bar{\alpha}_f\}$.

Theorem 2 (Query-level FWER over structural facets)**.** *Fix an episode contract, a query q, and $\alpha_{query} \in (0,1)$. Set $\alpha_f = \alpha_{query}/|F_{struct}(q)|$ for $f \in F_{struct}(q)$ and $\bar{\alpha}_f = \alpha_f/T_f$ (Eq. 2). If Theorem 1 holds for every shortlisted null pair, then the probability that Safe-Cover emits at least one false* structural *facet-support certificate within the query is at most α_{query}. Optional facets ($\bar{\alpha}_f = 1$) do not enter the bound. Proof in Appendix R.*

Certificate semantics. Theorem 2 controls the probability of *at least one false structural declaration within the query's tested family under the logged contract*. A passing pair is a partial structural declaration; complete certification requires every mined structural facet to pass. The guarantee is conditional on the contract-defined null and within-bin exchangeability assumptions.

The guarantee is conditional on replaying the same frozen shortlisting policy and on calibration/test nulls being exchangeable within the Mondrian bin. It can fail under policy drift, adaptive threshold changes, changed facet mining, or null pools that do not match test-time shortlisting. In those cases TRIDENT should enter heuristic mode or abstain rather than emit certified support. A corpus-drift stress test confirms this contract boundary empirically (Appendix H): identity calibrators respect nominal Type-I, while mismatched 2Wiki/MuSiQue→HotpotQA calibrators violate τ=0.10 by 7.6–8.9 percentage points, motivating re-calibration when the corpus or shortlisting distribution changes.

Safe-Cover defines the certified cover objective

$$\min_{S \subseteq P} \sum_{p \in S} c(p) \quad \text{s.t.} \quad \bigcup_{p \in S} C(p) \supseteq F_{\text{struct}}(q), \tag{3}$$

and emits a certificate only if a feasible cover fits within B_{ctx}; otherwise it abstains. In implementation, we use a deterministic greedy set-cover heuristic that picks the affordable passage maximising $|C(p) \cap U|/c(p)$ at each step, with ties broken by the frozen shortlist order. Greedy affects realised cost and abstention, but not the false-certificate guarantee, which depends only on the calibrated passage–facet tests. For multi-hop bridge queries, Safe-Cover runs a two-pass cover with a Hop-1-conditioned shortlister; both passes are frozen in the contract, so Theorem 2 extends without modification. Provenance-bound extraction, abstention reason codes, handoff branches, and replay validity are in Appendix C; a worked Polish-Russian War trace appears in Appendix Q (Figure 7).

3.4 Pareto-Knapsack: the efficiency regime

Pareto-Knapsack is TRIDENT's efficiency-oriented operating regime. It uses the same frozen passage–facet table $T(q)$ as Safe-Cover, but disables certificate thresholding and selects passages by marginal facet gain per token. It optimizes facet coverage under a hard evidence-token budget. In the default Pareto-500 setting, the marginal facet gain is binary:

$$\Delta(p \mid S) \;=\; \left| \{f \in F(q) : u(p,f) > 0\} \;\setminus\; \bigcup_{p' \in S} \{f : u(p',f) > 0\} \right|, \tag{4}$$

and the greedy score is $\Delta(p \mid S)/c(p)$. Soft verifier-score utilities are evaluated only as ablations (Section 4.6). Pareto-Knapsack halts when no affordable passage adds positive marginal facet gain and applies the same provenance check (Algorithm 2, Appendix C).

Operating modes. Safe-Cover is a statistical procedure: its output is a structural-facet support certificate with query-level FWER control under the logged contract (Theorem 2). Pareto-Knapsack is an efficiency procedure: its output is a budgeted evidence pack optimized for QA quality per token. The former is evaluated by false-certificate control; the latter is evaluated by F1 and evidence-token cost in the main paper, with latency reported in Appendix B. A theorem-by-theorem boundary and the full mode contrast are in Appendix D (Table 10). Three downstream certificate tiers — *struct-cert* (≥ 1 structural certificate per query), *answer-linked* (final answer cites a certified passage), and *full-cover* (all mined structural facets certified) — are reported as empirical diagnostics; only the false-cert event from Theorem 2 carries a validity guarantee.

4 Experiments

We evaluate TRIDENT as a two-regime framework. The Safe-Cover experiments test whether replay calibration restores post-selection validity and whether structural-facet certificates remain controlled under deployed contracts. The Pareto-Knapsack experiments test whether the same frozen passage–facet table yields a stronger F1–token operating point when the deployment goal is efficient QA rather than certification. Concretely, we organize the experiments around five claims, in validity-first order. **Q1 (validity).** Does pre-selection calibration fail on shortlisted nulls, and does replayed calibration recover Type-I / FWER control? **Q2 (support-certificate validity).** Under rebuilt miner-specific conformal pools, do Safe-Cover false-certificate rates respect the query-level bound? **Q3 (end-to-end certificate usefulness).** Do answer-linked structural certificates identify higher-quality answer slices? **Q4 (token-efficient QA).** Does Pareto-Knapsack improve mean matched-budget multi-hop F1 over compression, reranking, and selector baselines? **Q5 (boundaries and ablations).** Which components determine certificate coverage, and where does the framework fail? Safe-Cover and Pareto-Knapsack are evaluated on different metric families (Table 10); we do not compare them on the same metric.

4.1 Setup

We evaluate on HotpotQA [Yang et al., 2018], 2WikiMultiHopQA [Ho et al., 2020], and MuSiQue [Trivedi et al., 2022] with Llama-3-8B-Instruct [AI@Meta, 2024] at temperature 0 (Qwen3-8B [Team, 2025] and a KILT-FEVER [Thorne et al., 2018, Petroni et al., 2021] single-claim boundary case in Appendix B). Retriever, reranker, and verifier are shared across regimes; differences come solely from selection logic. Pareto-Knapsack is run at evidence-token caps of $400/500/1500$; Safe-Cover is reported at $B_{\text{ctx}}{=}2000$. Safe-Cover end-to-end runs use $\alpha_{\text{query}}{=}0.10$ on two deployed LLM-miner contracts—lexical (LLM+lex) and NLI (LLM+NLI) shortlisting—each backed by a contract-rebuilt conformal pool; regex-miner contract, token accounting, and CIs are in Appendices K–E, and the full main-results table is Appendix Table 5.

4.2 Post-selection validity (Q1)

We test Proposition 1 directly: for each held-out test query, score shortlisted null pairs with a *pre-selection* calibrator (every train null) and a *selection-conditional* calibrator (only top-T_f train nulls per (query, facet)). At nominal $\tau{=}0.05$, pre-selection rejects $0.157/0.087/0.180$ on HotpotQA/2Wiki/MuSiQue (a 1.7–$3.6\times$ inflation over nominal) while selection-conditional rejects $0.047/0.035/0.045$ (Figure 3; per-τ and per-facet breakdowns in Appendix F). Composing per-pair rates with the Bonferroni allocation, empirical query-level FWER respects nominal α_{query} at every reported level (Appendix I). Ignoring the shortlister rejects far more than nominal; replaying it restores the bound. This validates the calibration object; the next sections ask whether the resulting certificates remain controlled under miner-specific contracts (§4.3) and are useful in end-to-end QA (§4.4).

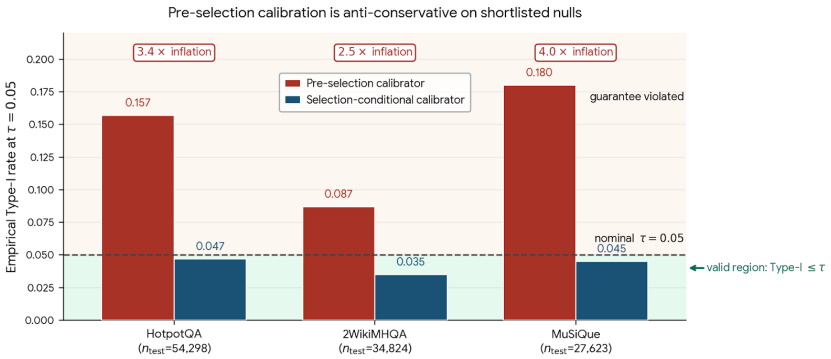

Figure 3: Empirical Type-I at $\tau{=}0.05$ on shortlisted nulls; only the selection-conditional calibrator stays near nominal.

Isolating replay from packing. Replay-conformal point estimates meet nominal in $540/540$ splits (108 calibration-only configurations). Replayed empirical quantiles nearly match this correction (Appendix S.2).

4.3 Safe-Cover validity and deployment diagnostics (Q2)

Table 1: **Safe-Cover audit and deployment diagnostics at $\alpha_{\text{query}}=0.10$.** Audit columns (false-cert / clean-cert / precision) validate Theorem 2; full-dev columns (answer-linked, abstention) carry no FWER claim. Column definitions and per-contract setup in §4.3.

Dataset	Contract	False-cert (%) ↓	Clean-cert (%) ↑	Prec. (%) ↑	Ans-link (%) ↑	Abst. (%) ↓
HotpotQA	LLM+lex	5.00 [3.41, 7.28]	1.40	21.9	6.94	4.65
HotpotQA	LLM+NLI	5.67 [4.08, 7.81]	2.83	33.3	8.40	4.67
2Wiki	LLM+lex	**1.60** [0.62, 4.04]	5.60	**77.8**	9.38	0.77
2Wiki	LLM+NLI	2.00 [0.86, 4.60]	5.60	73.7	**16.56**	0.87
MuSiQue	LLM+lex	6.60 [4.74, 9.12]	6.00	47.6	8.69	2.65
MuSiQue	LLM+NLI	8.33 [6.38, 10.82]	7.33	46.8	2.32	1.74

Swapping either the miner or the shortlister changes the selected hypothesis distribution, so we rebuild the conformal null pool separately for each deployed Safe-Cover contract on a disjoint calibration/audit split (details in Appendix K). Table 1 reports five per-contract quantities at $\alpha_{\text{query}}=0.10$. *False-cert* and *clean-cert* are audit-pool rates of ≥ 1 false / ≥ 1 valid structural certificate per query; *precision* is clean-cert / (clean-cert + false-cert); *answer-linked yield* and *deploy abstention* are full-dev deployment frequencies under the same frozen contract. The validity quantities and the deployment quantities are measured on different pools and are not directly comparable; we report them side-by-side to separate the statistical validity claim from the practical deployment yield.

All six false-cert point estimates lie below 0.10; MuSiQue LLM+NLI is closest (8.33%, CI upper bound 10.82%). The full-dev ledger contains at least one passing structural declaration on 2.73–16.82% of queries, with complete mined-facet certification on 0–0.59%. Appendix S.1 reports partial declaration coverage, retained certified passages, and complete cover separately.

4.4 End-to-end certificate usefulness (Q3)

Q2 checks contract-relative error rates; the deployment question is whether the certificates are also *useful*. We run full-dev Safe-Cover at $\alpha_{\text{query}}=0.10$ under the two deployed LLM contracts and compare per-query EM/F1 on the answer-linked slice (final answer cites a structurally certified passage) against the no-cert slice; bootstrap 95% CIs use $B=10{,}000$ resamples of the difference in means.

Across well-sized contracts, answer-linked certificates mark higher-quality answer slices: HotpotQA improves by ≈ 5 EM under both LLM shortlisters, 2Wiki LLM+NLI improves by $+16.63$ EM / $+14.15$ F1 (both CIs $\geq +11.95$), and MuSiQue LLM+lex improves by $+9.24$ EM (CI strictly above zero). MuSiQue LLM+NLI yields too small a slice ($n_{\text{link}}=56$) for a CI that excludes zero. These slice differences are observational associations: certified queries may be intrinsically easier. Full slice table with bootstrap CIs and the regex-miner parity row are in Appendix L (Table 23).

4.5 Efficiency regime: Pareto-Knapsack for low-token multi-hop QA (Q4)

After evaluating TRIDENT's certification regime, we test whether the same representation also improves efficient QA. Pareto-Knapsack disables certificate thresholding and uses $T(q)$ as a facet-aware utility table, selecting evidence by marginal structural gain per token. We evaluate this mode by matched-budget QA quality and evidence-token cost (latency results in Appendix B).

At a 500-token cap, Pareto-Knapsack obtains MH-F1 41.74 versus RECOMP's 40.98 with 16.3% fewer mean evidence tokens (Table 2). RECOMP wins on HotpotQA and 2Wiki; TRIDENT wins on MuSiQue. Appendix S.5 gives the RECOMP protocol and the native Search-R1/HippoRAG2 settings. Facet-aware packing drives Pareto's QA gains (§4.6).

Table 2: **500-token selector configurations on multi-hop QA.** F1 (%); MH-F1 is the mean over HotpotQA/2Wiki/MuSiQue; EvTok is mean evidence tokens. RECOMP follows the shared protocol; learned-compressor settings are in Appendix S.5.

Method	HotpotQA	2Wiki	MuSiQue	MH-F1	EvTok
TRIDENT Pareto-500	58.22	36.25	**30.74**	**41.74**	408
RECOMP-500	**59.87**	**41.37**	21.70	40.98	487.52
Sentence-Compress-500	51.29	36.37	21.75	36.47	495
LLMLingua-500	49.82	33.49	18.29	33.87	532
BGE-Rerank-500	46.05	30.96	27.11	34.71	470
LLM-Select-500	41.69	33.87	10.44	28.67	220
NLI-Filter-500	39.71	21.65	13.48	24.95	448

4.6 Boundaries and ablations (Q5)

Certificate power. Safe-Cover can certify only structural facets emitted by the miner and only passage–facet pairs that the verifier separates from shortlisted nulls. Low clean-cert rates therefore expose a verifier/miner bottleneck rather than a violation of the statistical guarantee. This is clearest on HotpotQA, where bridge-supporting passages are often close to bridge-adjacent nulls under the current verifier. Improving the entailment scorer, using passage-neighborhood support tests, or training a verifier on shortlisted hard nulls should increase certificate yield without changing the post-selection calibration principle.

Model-assisted facet–hop alignment and a rebuilt second-verifier contract characterize miner coverage and verifier sensitivity (Appendix S.3).

Contract dependence. Best certificate yield depends on the deployed contract: NLI shortlisting raises answer-linked volume on HotpotQA and 2Wiki, while lexical shortlisting is stronger on MuSiQue. Symbolic miners can be high-precision where support structure matches (regex on 2Wiki: small but high-quality slice; Appendix K).

Changes to calibration-critical components require rebuilding before certified emission resumes; rebuilt alternative-verifier and adaptive contracts have below-nominal point estimates (Appendices S.2 and S.3); rebuild cost and drift are in Appendix S.4.

Mechanism ablations. Table 3 separates what drives Pareto F1 from what drives Safe-Cover validity (full matrix in Tables 7, 8, Appendix B.1). Removing facet cover causes the largest F1 drop, with No-Cover losing 13.3 F1 on HotpotQA and 16.9 F1 on MuSiQue and Threshold-Then-Pack dropping further while abstention spikes. In contrast, replacing the conformal p-value with the raw NLI score (Raw-NLI utility) tracks Conformal Pareto within ≤ 1.7 F1 on the matched n=1,000 subset: selection-conditional calibration is Safe-Cover's validity-recovery ingredient, *not* Pareto's F1 driver. Miner quality is the dominant ceiling on certificate yield. The LLM-planner study is not an end-to-end comparison to the regex miner: it scores LLM-emitted facets against gold-supporting passages, so it isolates the structural-facet emission ceiling rather than the deployment pipeline. We therefore report the 32–177$\times$ gap at α_{query}=0.10 as an offline gold-support ceiling showing that structural facet emission, rather than the replay-calibration machinery, is the main yield bottleneck; the rigorous deployment audit under rebuilt LLM-miner contracts is Table 1. The choice of multiple-testing rule (per-facet Bonferroni, flat Bonferroni, Holm) is not a yield driver, because uniform T_f makes them coincide (Appendix K).

Table 3: **Mechanism ablations.** F1 (%) at the 500-token cap, Llama-3-8B-Instruct. *Abst. H/M* is abstention (%) on HotpotQA/MuSiQue. The *calibration-attribution* block uses a matched n=1,000 subset (seed 42) and is not row-comparable to the *selection-mechanism* block.

Block	Config	HotpotQA F1	2Wiki F1	MuSiQue F1	Abst. H/M (%)
Selection mechanism (full split)	TRIDENT Pareto-500	**58.22**	**36.25**	**30.74**	1.5 / 6.0
Selection mechanism (full split)	− facet cover (No-Cover)	44.92	21.15	13.79	0.0 / 0.0
Selection mechanism (full split)	− cover, + threshold filter	36.81	22.75	10.50	**7.9 / 18.8**
Calibration attribution (n=1,000)	Conformal Pareto	55.12	31.49	17.78	2.6 / 9.4
Calibration attribution (n=1,000)	− calibration (Raw-NLI utility)	53.44	32.03	18.46	2.6 / 9.4

5 Conclusion

TRIDENT recasts RAG support verification as selected-hypothesis testing: a frozen episode contract induces a shared passage–facet table that Safe-Cover reads as a calibrated testing table (query-level FWER-controlled structural declarations) and Pareto-Knapsack reads as a utility table (token-efficient multi-hop QA). The recipe—freeze the selection policy, replay it during calibration, compose tests at the query level—extends under the stated assumptions to retrieve-then-verify pipelines in which pre-selection thresholds would otherwise silently inflate Type-I error.

Acknowledgments and Disclosure of Funding

This work was supported in part by National Key R&D Program of China No. 2024YFB2705300, NSFC Grant 62232011, 62402315, the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX026D09), and State Key Laboratory of Submarine Geoscience (SGLabZZKT2025-07).

The authors declare no competing interests.

References

AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024.

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. URL http://arxiv.org/abs/2107.07511. arXiv:2107.07511 [cs].

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*, 2024.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2), April 2013. ISSN 0090-5364. doi: 10.1214/12-AOS1077. URL http://arxiv.org/abs/1306.1059. arXiv:1306.1059 [math].

Jiale Deng, Yanyan Shen, Ziyuan Pei, Youmin Chen, and Linpeng Huang. Influence Guided Context Selection for Effective Retrieval-Augmented Generation, October 2025. URL http://arxiv.org/abs/2509.21359. arXiv:2509.21359 [cs] TLDR: Inspired by recent advances in data selection, this novel metric quantifies context quality by measuring the performance degradation when removing each context from the list, effectively integrating query-aware relevance, list-aware uniqueness, and generator-aware alignment.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal Inference After Model Selection, April 2017. URL http://arxiv.org/abs/1410.2597. arXiv:1410.2597 [math].

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *International Conference on Learning Representations*, 2023.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Gautier Izacard and Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning, August 2025. URL http://arxiv.org/abs/2503.09516. arXiv:2503.09516 [cs].

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), June 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1371. URL http://arxiv.org/abs/1311.6238. arXiv:1311.6238 [math].

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*, volume 33, 2020.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. GraphReader: Building Graph-based Agent to Enhance Long-Context Abilities of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024a.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. AttributionBench: How Hard is Automatic Attribution Evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024b.

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6599–6612, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.334. URL https://aclanthology.org/2025.emnlp-main.334/.

Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. FIT-RAG: Black-Box RAG with Factual Information and Token Reduction. *ACM Transactions on Information Systems*, 43(2):1–27, March 2025. ISSN 1046-8188, 1558-2868. doi: 10.1145/3676957. URL https://dl.acm.org/doi/10.1145/3676957. 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. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Artyom Sorokin, Nazar Buzun, Alexander Anokhin, Egor Vedernikov, Petr Anokhin, Mikhail Burtsev, and Evgeny Burnaev. Q-RAG: Long Context Multi-Step Retrieval via Value-Based Embedder Training. 2026.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

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. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Jingjin Wang and Jiawei Han. PropRAG: Guiding Retrieval with Beam Search over Proposition Paths. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6223–6238, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.317. URL https://aclanthology.org/2025.emnlp-main.317/.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to Filter Context for Retrieval-Augmented Generation, November 2023. URL http://arxiv.org/abs/2311.08377. arXiv:2311.08377 [cs].

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. InfoGain-RAG: Boosting Retrieval-Augmented Generation through Document Information Gain-based Reranking and Filtering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7201–7215, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.365. URL https://aclanthology.org/2025.emnlp-main.365/.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed Resources For General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 641–649, New York, NY, USA, July 2024. Association for Computing Machinery. ISBN 979-8-4007-0431-4. doi: 10.1145/3626772.3657878. URL https://dl.acm.org/doi/10.1145/3626772.3657878.

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. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *International Conference on Learning Representations*, 2021.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. In *International Conference on Learning Representations*, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective Retrieval Augmented Generation, October 2024. URL `http://arxiv.org/abs/2401.15884`. arXiv:2401.15884 [cs].

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL `https://aclanthology.org/D18-1259/`.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. End-to-End Beam Retrieval for Multi-Hop Question Answering. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1718–1731, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.96. URL `https://aclanthology.org/2024.naacl-long.96/`.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. Multi-Step Reasoning Over Unstructured Text with Beam Dense Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

A Limitations

We collect the assumptions on which the paper's claims rest, and the regimes in which they are known to bind, so the contribution can be assessed at the right granularity. Table 4 summarises the dominant failure modes, where each is observed in our experiments, and the natural remedy.

Table 4: Dominant TRIDENT failure modes, where they are observed, and the natural remedy for each.

Failure mode	**Where observed**	**Likely remedy**
Missing structural facet	Regex miner; HotpotQA / 2Wiki comparisons	LLM facet planner
No covering passage	MuSiQue grounded-no-cert tier ($\sim$28%)	Stronger retriever / verifier
Diffuse support	MuSiQue multi-hop evidence	Pair-level / neighborhood verifier
Budget infeasible	Long evidence contexts	Compression before cover
p-value infeasible	Small α_{query} or sparse bins	Larger null pools or bin merging
Extraction failure	FEVER / open generation	Stronger provenance-bound extractor

Contract-scoped validity. Safe-Cover certificates are claimed only under the logged episode contract: the same retriever snapshot, deterministic shortlisting rules (including tie-breaks and T_f), verifier weights, Mondrian binning key, calibration pools, evidence serialization, structural-type set, abstention-handoff flags, and tie-break seed must be replayed at test time. If a calibration-critical element changes, the validity claim does not transfer; theorem-backed emission must pause until rebuilding and audit under the updated contract (Appendix S.4).

Calibration exchangeability and distribution shift. Theorem 1 requires $s(p, f)$ to be exchangeable with the bin-b calibration scores conditional on shortlisting. We engineer for this by replaying the identical shortlister against held-out negatives, verify the per-pair Type-I rate (Appendix G), and verify the query-level FWER (Appendix I). To quantify what happens when exchangeability

is broken, Appendix H runs a corpus-drift stress test: build the calibrator on dataset A and apply it to dataset B's held-out null pool. Within-dataset (identity) calibrators respect the bound at every $\tau \in \{0.01, 0.05, 0.10\}$. The HotpotQA-trained calibrator transfers cleanly to 2Wiki and MuSiQue. *The 2Wiki and MuSiQue calibrators applied to HotpotQA pairs violate the bound by 7.6–8.9 percentage points at* $\tau = 0.10$. The deployment lesson is what the contract already encodes: re-calibrate when the corpus changes; version checks trigger rebuilding under the updated contract (Appendix S.4).

Bonferroni allocation is not the binding conservatism. Replacing per-facet Bonferroni with Holm step-down at the query level produces *identical* empirical FWER on the held-out null pool and identical support-certificate rates on the LLM-miner ablation (Appendix K). When T_f is uniform the per-facet Bonferroni, flat Bonferroni, and Holm step-1 thresholds all coincide. The conservatism lives in the per-pair p-value distribution (mean $\pi > 0.97$ on BRIDGE_HOP and RELATION), not in the union bound.

Facet miner and verifier jointly bound certified coverage. The system can certify only what it can express as facets, and only what the verifier can discriminate on null passages. The guarantee is conditional on $F_{\text{struct}}(q)$, not on q as posed. On a 1,000-query random sample, the regex miner emits zero BRIDGE-HOP facets per query and recovers bridge structure via two-pass binding (Appendix J); 87.7% of bridge anchors and 74.4% of comparison anchors do land in gold supporting passages. Replacing the regex with a Llama-3-8B-Instruct facet planner and rebuilding the LLM-miner conformal null pool separately for each deployed shortlister (Section 4.3, Table 1) keeps the empirical query-level false-certificate point estimate below the nominal α_{query}=0.10 on every multi-hop dataset and every (LLM+lex, LLM+NLI) contract, and produces an answer-linked slice with substantial EM/F1 lift over the no-cert slice on every well-sized contract (Table 23). HotpotQA's clean-cert rate stays low (1.4–2.8%) because the off-the-shelf NLI verifier discriminates LLM-emitted BRIDGE_HOP hypotheses against bridge-adjacent null passages less reliably than ENTITY-anchored regex pairs. The structural-FWER-carrying slice is therefore *verifier-bound* on bridge-style entailment; a stronger entailment scorer or a pair-/neighborhood-level verifier is the natural remediation.

Verifier quality controls certificate feasibility. Conformal p-values inherit the verifier's discrimination. Poor discrimination can produce large p-values and few passing declarations; a sparse calibration pool separately limits rank resolution. Rebuilt second-verifier contracts yield below-nominal false-event point estimates, while declaration membership changes substantially (Appendix S.3). These results characterize verifier-specific validity and declaration membership.

Single-passage Safe-Cover is conservative on diffuse evidence. Safe-Cover certifies via a single winning passage. On MuSiQue, support for some structural facets is genuinely distributed and no single passage clears $\bar{\alpha}_f$. The system fails closed (`NO_COVER`). Pair-level or neighborhood entailment is a natural fix and is orthogonal to the contract machinery.

Pareto-Knapsack does not certify, and conformal calibration is not what drives Pareto's F1. The Pareto regime drops per-query statistical claims by design. We attribute Pareto's empirical gain by swapping the conformal p-value for the raw NLI score and re-running Pareto-500 on a matched n=1,000 subset (Table 8, *Calibration attribution*). The raw-NLI variant tracks conformal Pareto within ≤ 1.7 F1 on every dataset and is in fact slightly stronger on 2Wiki and MuSiQue. Selection-conditional calibration is therefore the validity-recovery ingredient that Safe-Cover requires (Section 4.2); it is not the load-bearing component of Pareto's matched-budget F1. Pareto's gains come from facet-aware coverage (No-Cover and Threshold-Then-Pack drop 13–21 F1 on HotpotQA at the same budget). The calibration ablation isolates this contribution at matched sample size.

Selection baselines. Appendix M compares against LLMLingua, Sentence-Compress, BGE-Rerank, LLM-Select, LLM-Select+NLI, and NLI-Filter. Appendix S.5 reports full-split RECOMP at 500 tokens and the native Search-R1/HippoRAG2 settings.

Scope of evaluation. We focus on multi-hop QA (HotpotQA, 2Wiki, MuSiQue) with KILT-FEVER as a single-claim boundary case. Broader tasks (summarization, long-form synthesis, code grounding) may require different facet designs, verification signals, and cost models; the contract machinery transfers, but coverage and verifier choice are task-dependent.

B Full main-paper tables and quality–cost summary

This appendix collects the full versions of tables that appear compressed in the main paper: the full main-results table covering both LLMs (Table 5), the full baseline comparison including agentic baselines (Table 6), and the full ablation matrix (Tables 7, 8). Figure 4 reports the headline quality–cost frontier; per-facet pre-selection vs. selection-conditional Type-I rates are in Appendix F.

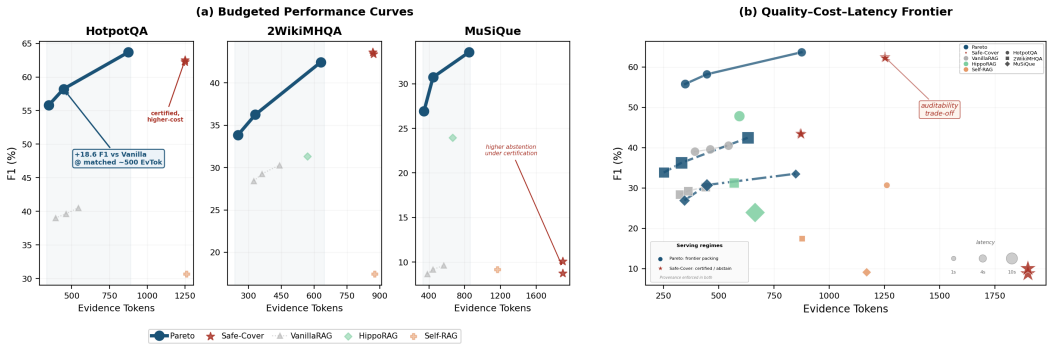

Figure 4: TRIDENT as a quality–cost system under hard evidence budgets. **(a)** Pareto traces monotone budget–quality curves and beats matched-budget truncation baselines; at $\sim$500 evidence tokens on HotpotQA, Pareto improves F1 from 39.61 to 58.22 over VanillaRAG. **(b)** Quality–cost–latency operating points. Point size indicates mean latency; provenance is enforced in both regimes. Inset: KILT-FEVER against IRCoT-Llama using total input tokens.

Table 5: TRIDENT main results. Each column pair is EM / F1 (%); MH-F1 is the mean F1 over HotpotQA/2Wiki/MuSiQue. EvTok and Lat are averaged over datasets with the corresponding accounting. The Qwen3-8B panel holds retrieval/verifier/calibration/cap fixed and changes only the LLM. KILT-FEVER EM is raw (‡); label-map normalized EM is shown in parentheses where it differs (†).

	2Wiki		HotpotQA		MuSiQue		FEVER‡		MH-F1	EvTok	Lat
Method	EM	F1	EM	F1	EM	F1	EM	F1	**(avg)**	(avg)	(ms)
Llama-3-8B-Instruct: Pareto-Knapsack (evidence-budget sweep)											
Pareto-400	25.95	33.83	43.80	55.80	15.02	26.93	49.40	49.50	38.85	315	2,493
Pareto-500	28.16	36.25	45.30	58.22	19.21	30.74	54.80	54.90	41.74	408	3,054
Pareto-1500	33.81	42.45	50.23	63.68	20.16	33.54	**77.50**	**77.70**	46.56	785	2,660
Llama-3-8B-Instruct: Safe-Cover (support certificates)											
Safe-2000 (equal)	41.48	48.93	45.64	58.90	8.32	14.39	57.36	57.38	40.74	567	4,183
Safe-4000 (loose)	41.59	48.96	45.56	58.92	8.56	14.67	57.35	57.37	40.85	565	4,212
Qwen3-8B (retrieval/verifier/calibration/cap fixed)											
Pareto-500	13.15	17.41	28.11	36.13	3.64	6.40	58.07‡	58.07	19.98	410	18,549
Pareto-1500	36.93	43.06	44.14	56.12	7.29	10.94	69.66‡ (61.60†)	69.66	36.71	784	25,013
Safe-2000 (equal)	**62.74**	**70.90**	**52.11**	**66.09**	**7.55**	**11.91**	**75.64**‡ (**76.18**†)	**75.65**	**49.63**	1,297	18,242

B.1 Ablations and learned-retriever positioning

Learned multi-step retrievers. Learned multi-step retrievers [Sorokin et al., 2026, Zhang et al., 2024, Jin et al., 2025, Li et al., 2024a] optimize fact recall under different training, LLM, and evidence-budget assumptions. They can supply TRIDENT's upstream candidate pool; changing the retriever requires replay calibration under the updated contract before Safe-Cover certifies its selected evidence.

Audit→end-to-end gap is a lexical-shortlister bottleneck. The rebuilt-pool audit predicts $1.4/6.4/6.6\%$ structural-cert coverage on HotpotQA/2Wiki/MuSiQue. To localize the gap we ran a controlled diagnostic on n=100 sampled queries per dataset: replay the same NLI verifier and

Table 6: System-level baseline comparison. Each cell is EM / F1 (%); EvTok is evidence tokens for budgeted methods and italicized total LLM-input tokens for agentic baselines that lack a budget cap. Search-R1 uses its released Qwen2.5-7B-Instruct PPO checkpoint (HotpotQA in-distribution, †); its FEVER entry uses label-mapped EM at 26.1% coverage§.

	2Wiki		HotpotQA		MuSiQue		FEVER		MH-F1	EvTok	Lat
Method	EM	F1	EM	F1	EM	F1	EM	F1	**(avg)**		(ms)
Llama-3-8B-Instruct comparisons											
TRIDENT Pareto-500	28.16	36.25	**45.30**	**58.22**	**19.21**	**30.74**	54.80	54.90	41.74	408	3,054
VanillaRAG-500	26.13	29.26	30.81	39.61	5.13	9.16	84.63	84.63	26.01	422	2,719
TRIDENT Pareto-1500	**33.81**	**42.45**	**50.23**	**63.68**	**20.16**	**33.54**	77.50	77.70	**46.56**	785	2,660
VanillaRAG-1000	27.05	30.28	31.57	40.55	5.83	9.62	86.64	86.64	26.82	517	2,679
HippoRAG2	27.42	31.32	37.79	47.82	15.35	23.95	**88.33**	**88.33**	34.36	608	5,176
Qwen3-8B comparisons											
TRIDENT Pareto-500	13.15	17.41	28.11	36.13	3.64	6.40	58.07	58.07	19.98	410	18,549
VanillaRAG-500	20.27	21.98	27.08	33.77	3.14	5.21	84.38	84.38	20.32	422	9,028
TRIDENT Pareto-1500	**36.93**	**43.06**	**44.14**	**56.12**	7.29	10.94	69.66	69.66	**36.71**	784	25,013
VanillaRAG-1000	21.32	23.23	27.87	34.70	3.14	5.35	88.59	88.59	21.09	517	9,164
HippoRAG2	21.14	23.75	34.77	44.01	14.77	20.47	**92.21**	**92.21**	29.41	727	24,145
Agentic / iterative baselines (unconstrained evidence budget)											
IRCoT (Llama-3-8B)	41.00	50.00	50.00	66.00	25.00	36.00	77.84	77.94	50.67	*4,909*	1,480
IRCoT (Qwen3-8B)	43.00	52.00	55.00	69.00	28.00	39.00	86.75	86.80	53.33	*5,753*	4,437
Search-R1 (Qwen2.5-7B, PPO)	57.00	65.00	57.05†	70.90†	2.00	6.00	24.49§	24.49§	47.30	*4,978*	2,134

Table 7: Ablation map. Each row pairs a claim with the ablation that targets it and the diagnostic the ablation should produce if the claim is true.

Claim	**Ablation**	**Expected diagnostic**
Calibration must be post-selection	Pre-selection calibrator	Inflated Type-I on shortlisted nulls
Facet coverage matters for Pareto F1	No-Cover / Threshold-Then-Pack	Lower F1 at matched token budgets
Calibration recovers validity, not Pareto F1	Raw-NLI Pareto vs. conformal Pareto	Similar F1, no per-query validity claim
Miner limits certification, not framework	Regex miner vs. LLM facet planner	Large lift in $\rho_{\text{cert}}^{\text{struct}}$
Budget controls token spend, not α	Threshold-α sweep	Flat EvTok across α, abstention varies
Quality–cost frontier is monotone	Pareto-400 / 500 / 1500 sweep	Monotone budget–quality curve

the same LLM-miner-rebuilt calibrator, but rank passages by NLI score (not the production lexical shortlister) and keep the top-T_f=4. Under that contract, structural-cert rates rise to $\mathbf{8\%/21\%/31\%}$ with gold passages surviving the shortlist for $89\%/84\%/79\%$ of queries; gold-passage median $\min p$ is $0.060/0.019/0.057$ vs. non-gold median ~ 0.99—the calibrator separates gold from null cleanly. The remediation is a facet-aware semantic shortlister, orthogonal to the calibrator and the FWER bound.

C Framework operational details

This appendix specifies the episode contract, facet instantiation, evidence costs, selection algorithms, and replay procedure.

Episode-contract fields and their roles.

Placeholder facets and instantiation. Multi-hop queries may introduce placeholder facets whose hypotheses are not meaningful until a binding value is available (e.g., BRIDGE-HOP2 carries a $\langle X \rangle$ slot). To avoid scoring ill-formed hypotheses, placeholder facets are skipped in the initial pass and evaluated only after instantiation by the Hop-1 binding extractor.

Candidate cost model. Each passage p has nonnegative cost $c(p)$ counting only the tokens contributed by p to the serialized evidence prompt. We impose a hard B_{ctx} and require $\sum_{p \in S} c(p) \leq B_{\text{ctx}}$. Generation tokens are tracked separately. The cost model is part of the contract.

Mondrian binning, bin merging, and infeasibility. Conformal p-values have a floor $1/(|\mathcal{N}_b| + 1)$. If the target threshold falls below this floor, certification is mathematically impossible. A fixed

Table 8: Full TRIDENT ablations. *Selection mechanism* replaces facet cover with *No-Cover* (pack by raw utility) and *Threshold-Then-Pack* (p-value pre-filter then pack). *Calibration attribution* swaps the selection-conditional conformal p-value for the raw NLI score in Pareto-Knapsack (no Mondrian binning, no calibration pool). Both rows of the calibration block use n=1,000 random-seeded subsets (seed 42). *Utility shape* compares binary marginal gain with a soft-sigmoid utility (`utility_tau`=0.05) and a no-rerank variant. *LLM interface* replays Safe-2000 with raw formatting instead of the chat template.

Config	Dataset	EM / F1	Abst.%	EvTok	TotTok	Lat p50
Selection mechanism (Pareto-500, Llama-3-8B; full data)						
No-Cover	HotpotQA	33.95 / 44.92	0.01	446	593	247
No-Cover	MuSiQue	7.65 / 13.79	0.00	465	636	553
Threshold-Then-Pack	HotpotQA	27.27 / 36.81	7.90	319	446	270
Threshold-Then-Pack	MuSiQue	5.54 / 10.50	18.78	252	378	188
Calibration attribution (Pareto-500, Llama-3-8B; matched n=1,000, seed 42)						
Conformal Pareto (parity)	HotpotQA	44.08 / **55.12**	2.60	436	1,476	2,567
Raw-NLI utility	HotpotQA	42.47 / 53.44	2.60	435	1,473	2,857
Conformal Pareto (parity)	2Wiki	25.05 / 31.49	2.60	328	1,235	2,627
Raw-NLI utility	2Wiki	26.01 / **32.03**	2.60	333	1,244	2,830
Conformal Pareto (parity)	MuSiQue	13.15 / 17.78	9.40	395	1,315	2,387
Raw-NLI utility	MuSiQue	12.60 / **18.46**	9.40	396	1,321	2,406
Utility shape (Pareto-500, α=0.6; full data)						
No-Rerank	HotpotQA	45.94 / 58.08	1.54	451	1,012	2,275
Soft Sigmoid	HotpotQA	46.14 / 59.20	1.54	467	1,025	531
Soft Sigmoid	MuSiQue	17.28 / 28.62	6.04	460	934	1,504
LLM interface (Safe-2000)						
No-Chat-Template	MuSiQue	5.13 / 10.26	8.77	845	—	19,685

Table 9: Episode contract fields frozen by TRIDENT. Freezing these fields makes the selected hypotheses and their support tests replayable bit-for-bit (Theorem 3).

Contract field	Why it is frozen
Retriever snapshot	Fixes the candidate pool P
Candidate cap and shortlist rule	Fixes the shortlisted passage–facet hypotheses
Verifier version	Fixes the score distribution $s(p, f)$
Mondrian binning key	Fixes calibration strata
Calibration pools $\mathcal{N}_b$	Fixes conformal p-values
Evidence serializer / tokenizer	Fixes evidence cost $c(p)$
Structural-type set $\mathcal{T}_{\text{struct}}$	Fixes the Bonferroni denominator
Budget B_{ctx}	Fixes feasible evidence sets
Abstention / handoff flags	Fixes typed reason codes and F1/F2/F3 routing
Tie-break seed	Enables bit-identical replay

bin-merging rule reaches a minimum effective pool size; merge depth is logged. If feasibility still cannot be met, the Safe-Cover returns `INFEASIBLE_PVALUE` rather than guess.

Heuristic-mode floor. For deployments without a Mondrian-calibrated negative pool, raw scores replace conformal p-values and the FWER guarantee of Theorem 2 is voided. To keep coverage feasible, a fixed lower bound α_{floor} is applied to the per-pair threshold of structural facets:

$$\bar{\alpha}_f^{\text{heur}} = \max(\bar{\alpha}_f, \alpha_{\text{floor}}), \qquad f \in F_{\text{struct}}(q). \tag{5}$$

Heuristic-mode runs are reported separately and do not appear in any support-certificate statistic.

Two-pass cover. For queries whose Hop-2 facets reference a value bound by Hop-1, Safe-Cover runs (1) a Hop-1 cover, (2) a constrained span extractor on $\text{Cert}_1(q)$ to recover the bridge entity $\hat{x}$, (3)

hypothesis instantiation with $\hat{x}$, (4) a Hop-2-conditioned candidate generation, and (5) a Hop-2 cover on the conditioned pool. Both passes share the same contract; only the per-facet shortlist changes. Theorem 2 extends without modification because the union bound still ranges over $F_{\text{struct}}(q)$.

Provenance-bound answer extractor (four tiers). An extracted output carries a support certificate iff its citing passage lies in $\mathrm{Cert}(q)$. Tiers, strict to permissive, with later tiers invoked only on NONE: (1) facet-targeted span extractor on the winning passage of an answer-bearing facet; (2) frozen-span extractor returning a verbatim span from any passage in $\mathrm{Cert}(q)$; (3) constrained generative extraction conditioned on $\mathrm{Cert}(q)$ alone with a question-type-aware prompt (wh, yes/no, FEVER SUPPORTS/REFUTES); (4) cited-passage upgrade for free-form generation, support-certified iff the cited passage lies in $\mathrm{Cert}(q)$.

Abstention reason codes. `NO_FACETS` ($F(q) = \emptyset$); `NO_PASSAGES` ($P = \emptyset$); `NO_SUPPORT_EDGES` (no (p, f) with valid π); `NO_COVER` (some structural facet uncovered); `INFEASIBLE_BUDGET` (cover exists but cannot fit the budget; dual lower bound on remaining cover cost is computed over $F_{\text{struct}}(q)$ only); `INFEASIBLE_PVALUE` (bin-floor infeasibility after merging); `HEURISTIC_BUDGET_INFEASIBLE` (analogue under Eq. 5).

Bonferroni allocation and fixed coverage sets. Eq. 2 induces a fixed coverage set $C(p) = \{f \in F(q) : \pi(p, f) \le \bar{\alpha}_f\}$. Optional facets ($\bar{\alpha}_f = 1$) appear in $C(p)$ but do not contribute to the bound. Algorithms 1 (Safe-Cover) and 2 (Pareto-Knapsack) state the deterministic greedy rules used in both regimes.

Algorithm 1 Safe-Cover (budgeted greedy facet cover)

Input: candidate pool P, costs $c(\cdot)$, coverage sets $C(\cdot)$, structural facets $F_{\text{struct}}(q)$, evidence budget B_{ctx}
$S \leftarrow \emptyset$; $\mathrm{Cert} \leftarrow \emptyset$
$U \leftarrow F_{\text{struct}}(q)$; $b \leftarrow B_{\text{ctx}}$
while $U \neq \emptyset$ **do**
  $\mathcal{F} \leftarrow \{p \in P \setminus S : c(p) \le b,\ C(p) \cap U \neq \emptyset\}$
  if $\mathcal{F} = \emptyset$ **then**
    return ABSTAIN(NoCover/InfeasibleBudget)
  end if
  $p^\star \leftarrow \arg\max_{p \in \mathcal{F}} |C(p) \cap U| / c(p)$
  $S \leftarrow S \cup \{p^\star\}$; $b \leftarrow b - c(p^\star)$
  emit $\{(p^\star, f) : f \in C(p^\star) \cap U\}$ into Cert
  $U \leftarrow U \setminus C(p^\star)$
end while
return (S, Cert)

Algorithm 2 Pareto-Knapsack (utility-per-cost greedy)

Input: pool P, costs $c(\cdot)$, utilities $u(p, f)$, budget B_{ctx}, with $U(S) = \sum_{f \in F(q)} \max_{p \in S} u(p, f)$
$S \leftarrow \emptyset$; $b \leftarrow B_{\text{ctx}}$
while $\mathcal{F} \leftarrow \{p \in P \setminus S : c(p) \le b\} \neq \emptyset$ **do**
  $\Delta U(p \mid S) \leftarrow U(S \cup \{p\}) - U(S)$
  $p^\star \leftarrow \arg\max_{p \in \mathcal{F}} \Delta U(p \mid S) / c(p)$
  if $\Delta U(p^\star \mid S) \le 0$ **then**
    break
  end if
  $S \leftarrow S \cup \{p^\star\}$; $b \leftarrow b - c(p^\star)$
end while
if provenance check on S fails **then**
  return ABSTAIN(NoProvenance)
end if
return S

Handoff branches. F1 hard-abstain returns the typed reason code; F2 Pareto-handoff (the production setting) passes control to the Pareto-Knapsack selector over the same candidate pool; F3 best-effort within Safe-Cover packs by $(\min_f \pi(p, f), c(p), \mathrm{pid})$ when greedy returns $|S| = 0$ and is excluded from any support-certificate statistic. A generative recovery layer may be invoked on the resulting set; recovery answers obey the same provenance-pass-through rule.

Replay validity.

Theorem 3 (Replay validity)**.** *Let $\mathcal{C}(q, E)$ denote the certificate payload (including any `ABSTAIN` reason code) returned by Safe-Cover for query q under contract E. If every element of E is replayed byte-identically, then $\mathcal{C}'(q, E) = \mathcal{C}(q, E)$ bit-for-bit. Proof in Appendix R.*

Diagnostics and drift fallback. The system logs certificate payloads, shortlist sizes, bin depths, near-threshold counts, abstention codes, structural-vs-optional facet counts, the chosen handoff branch, the per-query support-certificate indicator ρ_{cert}, and version metadata sufficient for replay. Per-bin score summaries and near-threshold counts are monitored as drift signals; if a drift alarm triggers, Safe-Cover 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 support-certificate FWER bound.

D Contribution–theorem–experiment map

Mode contrast. Table 10 contrasts Safe-Cover and Pareto-Knapsack along purpose, formal guarantee, selection rule, output, and main metric. Theorem 2 controls only the Safe-Cover false-cert event; Pareto's gains are empirical.

Table 10: TRIDENT's two operating modes share the same frozen passage–facet table $T(q)$ but answer different questions and carry different claims.

Claim	**Safe-Cover (validity regime)**	**Pareto-Knapsack (efficiency regime)**
Purpose	Certified evidence support	Budgeted QA deployment
Formal guarantee	FWER over false structural-facet certificates	No FWER claim; empirical quality–cost
Selection rule	Min-cost cover over certified facets	Utility-per-token greedy packing
Output	Certified evidence or typed abstention	Evidence pack + provenance check
Answer correctness	Not directly; certifies the support layer	Not directly; evaluated end-to-end
Main metric	Type-I / FWER / $\rho_{\text{cert}}^{\text{struct}}$	EM/F1 / EvTok / latency

Theorem-by-theorem boundary and contribution–experiment map. Table 11 records exactly which claims are covered by the theorems and which are empirical-only. Table 12 maps each contribution to its supporting theorem and experiment.

Table 11: What TRIDENT's theorems cover. Empirical-only claims appear in Section 4 and are not statements of validity.

Claim	**Covered by theorem?**
Marginal validity of p-values for shortlisted null pairs	Yes (Thm. 1)
Query-level Bonferroni FWER over false structural-facet certificates	Yes (Thm. 2)
Bit-identical Safe-Cover replay under contract	Yes (Thm. 3)
Optional/ENTITY certificate validity ($\bar{\alpha}_f = 1$)	No
Answer correctness	No
Facet-mining completeness	No
Generator faithfulness on selected evidence	No
Pareto-Knapsack per-query statistical validity	No
Pareto-Knapsack F1 improvement	Empirical only

E Evaluation protocol, statistical treatment, and latency definition

Confidence intervals. Table 13 provides 95% confidence intervals for systems and datasets with interval logs in Tables 5 and 6. These intervals quantify uncertainty for the original system comparisons with available interval logs.

VanillaRAG comparison. The Llama-3-8B Pareto-500 comparisons against VanillaRAG have non-overlapping intervals for both EM and F1 under the original matched setting (Table 13).

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 Lat50/Lat90/Lat95 under that policy.

Table 12: Claim map: contribution → theorem(s) → experiment(s).

Contribution	Theorem	Validating experiment
Pre-selection calibration is anti-conservative on selected hypotheses	Prop. 1	Empirically verified by replaying an unconditional-null calibrator end-to-end (Appendix G).
Episode-frozen conformal facet tests	Thm. 1	Per-pair Type-I at $\tau \in \{0.01, 0.05, 0.10\}$ on a held-out null pool (Table 15).
Budgeted min-cost cover with structural-FWER	Thm. 2	Query-level FWER on held-out null queries (Table 17); reliability curve (Fig. 5); mixed-pool Type-I/power (Table 18).
Replay validity of certificates	Thm. 3	Bit-identical certificate hashes under contract replay (operational; not a separate experiment).
Pareto improves mean matched-budget multi-hop F1	—	Tables 5, 6, 24; ablations Table 8.
Pareto's F1 gain attributes to facet cover, not to conformal calibration	—	*Calibration attribution* block of Table 8: raw-NLI vs. conformal Pareto-500 on matched n=1,000 subsets, ΔF1 $\leq$ 1.7 on every dataset; pair with the No-Cover / Threshold-Then-Pack rows that drop 13–21 F1 on HotpotQA.
Miner is the bottleneck (deterministic vs. LLM)	—	Tables 19, 20, 21, 22.

Compute environment. All experiments were run on a single local workstation equipped with $4\times$ NVIDIA RTX 4090 (24 GB) and $3\times$ NVIDIA RTX Pro 6000 (96 GB) GPUs. Individual end-to-end runs ranged from a few hours up to roughly 20 hours of wall-clock time depending on the dataset, LLM (Llama-3-8B-Instruct or Qwen3-8B), and selection regime (Pareto-Knapsack vs. Safe-Cover). The full project consumed additional compute beyond the runs reported in the paper for preliminary configurations, miner and shortlister sweeps, and conformal-pool calibration that did not appear in the final tables.

Token accounting. We distinguish (i) evidence tokens (EvTok), the number of tokens contributed by the evidence passages passed to the LLM, and (ii) total input tokens (TotTok), the full LLM input including the fixed prompt template and evidence. Both are computed using the LLM's tokenizer when available. FEVER Pareto runs log total input tokens but not a separate evidence-token split, so they are excluded from EvTok averages and reported instead in Table 24. Completion lengths are standardized via a shared decoding configuration across all methods.

F Pre-selection vs. selection-conditional Type-I (per-facet detail)

This appendix reports the per-facet breakdown of the pre-selection vs. selection-conditional Type-I experiment whose aggregate summary is plotted as Figure 3 in the main paper. We use a $50/50$ random split by query ID on the calibration set of each dataset, define the deterministic top-T ($T = 4$) shortlister within each (query, facet) group, and build two calibration pools from the train half: ***Pre*** (every train null score, unconditional on rank) and ***Sel*** (only the top-T train nulls per (query, facet)). For each test query we score its shortlisted (top-T) null pairs against both pools and record the empirical Type-I rejection rate at $\tau \in \{0.01, 0.05, 0.10\}$. Per Proposition 1, only the Sel column should respect nominal τ.

Reading the table. Pre rejects every shortlisted null at strictly higher rates than the nominal τ on every (dataset, facet) row where the verifier has discriminative power; the inflation is largest where shortlisting moves mass furthest from the unconditional null (ENTITY, NUMERIC). Sel respects nominal at every reported level. BRIDGE_HOP and RELATION appear conservative for both calibrators because the per-pair p-value distribution under the null is heavily skewed toward 1 on these typed facets (Appendix G); the inflation effect is real but small, since both calibrators' mass is at $\pi \approx 1$. The takeaway is the same on every well-sampled row: a calibrator that ignores the

Table 13: 95% confidence intervals for EM and F1 metrics when interval logs are available. Wilson score intervals for EM; stratified bootstrap (B=1000) for F1.

Method	Dataset	EM 95% CI	F1 95% CI
TRIDENT Pareto-500 (Llama-3-8B)			
Pareto-500	HotpotQA	[43.9, 46.9]	[56.9, 59.5]
Pareto-500	2Wiki	[26.9, 29.4]	[34.9, 37.5]
Pareto-500	MuSiQue	[15.5, 23.4]	[26.9, 34.7]
TRIDENT Pareto-1500 (Llama-3-8B)			
Pareto-1500	HotpotQA	[49.0, 51.6]	[62.6, 64.8]
Pareto-1500	2Wiki	[32.7, 34.8]	[41.4, 43.5]
Pareto-1500	MuSiQue	[16.9, 23.9]	[30.0, 37.4]
TRIDENT Safe-Cover (Llama-3-8B)			
Safe-2000	HotpotQA	[44.4, 46.8]	[57.9, 60.0]
Safe-2000	2Wiki	[40.6, 42.3]	[48.1, 49.7]
Safe-2000	MuSiQue	[7.2, 9.4]	[13.2, 15.6]
Safe-2000	KILT-FEVER	[56.4, 58.4]	[56.5, 58.4]
Safe-4000	HotpotQA	[44.4, 46.8]	[57.9, 60.1]
Safe-4000	2Wiki	[40.8, 42.5]	[48.2, 49.8]
Safe-4000	MuSiQue	[7.5, 9.7]	[13.5, 15.9]
Safe-4000	KILT-FEVER	[56.4, 58.3]	[56.4, 58.3]
TRIDENT Qwen3-8B			
Pareto-500	HotpotQA	[27.0, 29.2]	[35.0, 37.2]
Pareto-500	2Wiki	[12.6, 13.8]	[16.8, 18.1]
Pareto-500	MuSiQue	[2.8, 4.4]	[5.4, 7.3]
Pareto-1500	HotpotQA	[43.0, 45.3]	[55.1, 57.2]
Pareto-1500	2Wiki	[36.0, 37.8]	[42.2, 43.9]
Pareto-1500	MuSiQue	[6.2, 8.5]	[9.7, 12.2]
Safe-2000	HotpotQA	[51.0, 53.3]	[65.2, 67.1]
Safe-2000	2Wiki	[61.9, 63.6]	[70.1, 71.6]
Safe-2000	MuSiQue	[6.4, 8.7]	[10.7, 13.2]
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]
HippoRAG2	HotpotQA	[36.7, 38.9]	[46.6, 49.0]
HippoRAG2	2Wiki	[26.6, 28.3]	[30.3, 32.4]
HippoRAG2	MuSiQue	[13.7, 17.0]	[22.0, 25.9]

shortlister is not valid on the selected pool, and replaying the same shortlister at calibration time recovers marginal validity.

G Empirical certificate validity

Theorems 1 and 2 make a falsifiable prediction: under the logged episode contract, the conformal p-values for null pairs are super-uniform, and the rate of false structural-facet certificates respects the nominal bound. We empirically check this prediction directly. For each dataset and facet type, we replay the frozen contract on a held-out negative pool—$(passage, facet)$ pairs constructed from queries unseen at calibration time, with shortlisting and binning re-run under the contract—and record the conformal p-value for every shortlisted pair. The empirical Type-I rate at level τ is the fraction of pairs with $\pi(p, f) \leq \tau$; under Theorem 1 this should be at most τ. We also report a Kolmogorov–Smirnov statistic against $\mathrm{Uniform}[0, 1]$ and the mean/median p-value. The data below were collected on 2026-04-26.

Table 14: Per-facet pre-selection vs. selection-conditional Type-I on shortlisted held-out nulls ($T = 4$, seed 0). Pre is the rejection rate under a calibrator built on all train nulls; Sel is the rejection rate under the calibrator restricted to top-T train nulls per (query, facet). Inflation = Pre/Sel is the multiplicative anti-conservative factor. Facet rows with fewer than 200 test pairs are pooled into the ALL row to control sampling noise.

			$\tau = 0.01$		$\tau = 0.05$		$\tau = 0.10$	
Dataset	**Facet**	n	Pre / Sel	Infl.	Pre / Sel	Infl.	Pre / Sel	Infl.
HotpotQA	ALL	54,298	0.037 / 0.010	3.7×	0.157 / 0.047	3.4×	0.268 / 0.086	3.1×
HotpotQA	ENTITY	29,552	0.049 / 0.010	4.8×	0.225 / 0.051	4.4×	0.391 / 0.103	3.8×
HotpotQA	NUMERIC	7,944	0.033 / 0.013	2.6×	0.136 / 0.057	2.4×	0.252 / 0.105	2.4×
HotpotQA	RELATION	8,038	0.021 / 0.008	2.6×	0.040 / 0.040	1.0×	0.040 / 0.040	1.0×
HotpotQA	TEMPORAL	4,108	0.023 / 0.010	2.3×	0.107 / 0.048	2.2×	0.146 / 0.103	1.4×
HotpotQA	BRIDGE_HOP	4,560	0.010 / 0.008	1.2×	0.010 / 0.010	1.0×	0.010 / 0.010	1.0×
2Wiki	ALL	34,824	0.020 / 0.007	2.8×	0.087 / 0.035	2.5×	0.117 / 0.067	1.8×
2Wiki	ENTITY	19,988	0.031 / 0.010	3.1×	0.135 / 0.051	2.7×	0.176 / 0.099	1.8×
2Wiki	NUMERIC	3,208	0.017 / 0.008	2.2×	0.089 / 0.050	1.8×	0.165 / 0.095	1.7×
2Wiki	RELATION	6,004	0.003 / 0.003	1.0×	0.003 / 0.003	1.0×	0.003 / 0.003	1.0×
2Wiki	BRIDGE_HOP	4,280	0.000 / 0.000	—	0.000 / 0.000	—	0.000 / 0.000	—
MuSiQue	ALL	27,623	0.050 / 0.009	5.6×	0.180 / 0.045	4.0×	0.259 / 0.083	3.1×
MuSiQue	ENTITY	18,897	0.059 / 0.008	7.1×	0.228 / 0.051	4.5×	0.324 / 0.100	3.2×
MuSiQue	NUMERIC	1,816	0.041 / 0.003	12.5×	0.209 / 0.046	4.6×	0.368 / 0.093	4.0×
MuSiQue	RELATION	6,186	0.023 / 0.012	2.0×	0.023 / 0.023	1.0×	0.023 / 0.023	1.0×
MuSiQue	TEMPORAL	504	0.075 / 0.016	4.8×	0.240 / 0.077	3.1×	0.318 / 0.157	2.0×

The aggregate bound is empirically tight but conservative in practice. The population-level empirical Type-I rates are 0.0103, 0.0465, 0.0872 on HotpotQA; 0.0073, 0.0369, 0.0525 on 2Wiki; and 0.0093, 0.0419, 0.0830 on MuSiQue against nominal $\tau = 0.01, 0.05, 0.10$. Thus all aggregate rows are below nominal except the HotpotQA $\tau = 0.01$ row, which is essentially at the boundary (0.0103). This is the empirical analog of Theorem 1 on selection-conditional facet tests, evaluated directly under the frozen contract (i.e., on the same shortlister and binning that Safe-Cover uses at certification time). The query-level FWER bound of Theorem 2 composes these per-pair rates with Bonferroni allocation across $|F_{\text{struct}}(q)|$, so the aggregate per-pair conformity above is precisely the input the bound requires.

Conservatism varies by facet type. The KS statistic and the mean/median π columns reveal that the conformal p-values are typically super-uniform but *heavily skewed toward* 1, especially on BRIDGE_HOP (mean π above 0.97 where sampled) and RELATION (mean above 0.93). The verifier rarely produces low p-values on these structural facet types under the null—which is exactly why the FWER stays below nominal on most aggregate rows, and why typed conformal calibration is the right primitive for them. ENTITY and NUMERIC sit closer to but still above the diagonal (mean π in $[0.73, 0.92]$), and TEMPORAL is the most marginal of the well-sampled types. CAUSAL is outside the structural-type set covered by Theorem 2; on MuSiQue its p-values are near-uniform (KS $= 0.064$, mean $\pi = 0.50$) and its per-pair Type-I@0.10 is 0.153 ($n_{\text{test}} = 451$), while the MuSiQue aggregate remains within the bound.

Small-n facet types. The 2Wiki PROCEDURAL row ($n_{\text{test}} = 35$) shows Type-I@0.05 $= 0.171$, which is above nominal but on a sample where the binomial 95% upper bound at $\tau = 0.05$ is approximately 0.18, so it is consistent with the bound. The corresponding MuSiQue PROCEDURAL row ($n_{\text{test}} = 22$) reports zero false certificates. In deployment, bin-merging and minimum-pool rules already prevent these very-small-n facet types from emitting certificates without falling back to `INFEASIBLE_PVALUE`.

Empirical validation. The empirical bound is validated on the same selection contract used at test time, with aggregate rates below nominal except the near-bound HotpotQA $\tau = 0.01$ row. This supports the marginal-validity claim under selection (Theorem 1) and, by union bound, the query-level

Table 15: Empirical certificate validity by facet type. n_{test}: number of held-out shortlisted (p, f) pairs scored under the frozen contract. **KS**: Kolmogorov–Smirnov statistic against $\mathrm{Uniform}[0, 1]$ (large values indicate a conservatively skewed null distribution rather than a violation). **Type-I@τ**: empirical fraction of pairs with $\pi(p, f) \leq \tau$; the marginal-validity bound predicts this is $\leq \tau$. **Mean / median** π: aggregate p-value distribution. The aggregate ALL row per dataset is the directly comparable population-level Type-I rate; per-facet rows for very small n_{test} (CAUSAL, COMPARISON, PROCEDURAL) are reported for completeness but are subject to substantial sampling noise.

Dataset	Facet type	n_{test}	KS	Type-I@0.01	Type-I@0.05	Type-I@0.10	Mean π	Median π
2WikiMultiHopQA								
2Wiki	ENTITY	65,319	0.918	0.0100	0.0512	0.0702	0.922	0.989
2Wiki	RELATION	12,020	0.994	0.0013	0.0013	0.0013	0.994	0.995
2Wiki	BRIDGE_HOP	13,650	0.999	0.0000	0.0000	0.0000	0.999	0.999
2Wiki	NUMERIC	6,073	0.820	0.0071	0.0426	0.0939	0.838	0.993
2Wiki	TEMPORAL	1,777	0.645	0.0107	0.0152	0.0152	0.712	0.660
2Wiki	CAUSAL	623	0.121	0.0000	0.0225	0.0225	0.526	0.544
2Wiki	COMPARISON	5	0.800	0.0000	0.0000	0.0000	0.900	1.000
2Wiki	PROCEDURAL	35	0.229	0.0000	0.1714	0.2000	0.540	0.575
2Wiki	**ALL**	**99,502**	**0.908**	**0.0073**	**0.0369**	**0.0525**	—	—
MuSiQue								
MuSiQue	ENTITY	114,502	0.861	0.0102	0.0494	0.0997	0.871	0.971
MuSiQue	RELATION	27,805	0.977	0.0058	0.0058	0.0058	0.977	0.983
MuSiQue	NUMERIC	8,827	0.735	0.0087	0.0532	0.1071	0.769	0.922
MuSiQue	TEMPORAL	2,165	0.721	0.0106	0.0540	0.0905	0.757	0.811
MuSiQue	CAUSAL	451	0.064	0.0133	0.0998	0.1530	0.497	0.499
MuSiQue	BRIDGE_HOP	267	1.000	0.0000	0.0000	0.0000	1.000	1.000
MuSiQue	COMPARISON	51	0.926	0.0000	0.0000	0.0000	0.934	0.946
MuSiQue	PROCEDURAL	22	0.282	0.0000	0.0000	0.0000	0.561	0.600
MuSiQue	**ALL**	**154,090**	**0.816**	**0.0093**	**0.0419**	**0.0830**	—	—
HotpotQA								
HotpotQA	ENTITY	144,017	0.683	0.0106	0.0510	0.0996	0.733	0.938
HotpotQA	NUMERIC	23,126	0.805	0.0105	0.0501	0.0974	0.824	0.944
HotpotQA	TEMPORAL	10,030	0.813	0.0121	0.0566	0.0714	0.830	0.884
HotpotQA	RELATION	16,499	0.933	0.0095	0.0189	0.0189	0.935	0.952
HotpotQA	BRIDGE_HOP	9,074	0.978	0.0047	0.0047	0.0047	0.978	0.983
HotpotQA	COMPARISON	153	0.578	0.0000	0.0850	0.0850	0.670	0.662
HotpotQA	CAUSAL	45	0.215	0.0000	0.1111	0.1333	0.425	0.318
HotpotQA	**ALL**	**202,944**	**0.692**	**0.0103**	**0.0465**	**0.0872**	—	—

FWER bound (Theorem 2). Appendix J quantifies facet-miner robustness; Appendix S.4 evaluates contract rebuilding under distribution shift.

H Drift stress test

The paper claims contract-scoped validity: certificates carry the FWER bound only when the same retriever, calibrator, verifier, and serialization are replayed. We test the strongest assumption empirically by swapping the calibrator's training corpus while holding everything else fixed. For each ordered pair (calibrator dataset A, test dataset B), we (i) build the conformal pool from A's training half (50% of A's queries), (ii) score B's held-out null pool against that pool, and (iii) report empirical Type-I rate at $\tau \in \{0.01, 0.05, 0.10\}$.

Diagonals respect the bound. The within-dataset diagonals reproduce the result of Appendix G: empirical per-pair Type-I respects the nominal bound at all three τ levels.

Cross-corpus, the bound is not automatic. The 2Wiki and MuSiQue calibrators applied to HotpotQA test pairs violate the bound at every τ, by 1.0–8.9 percentage points: e.g., $\widehat{\mathrm{TypeI}}@0.10 = 0.189$ when calibrating on 2Wiki and testing on HotpotQA, against the nominal 0.10. This is exactly the failure mode the paper's "contract-scoped validity" clause predicts: the held-out negatives the calibrator was built on no longer represent the deployment-time null score distribution, so the rank statistic is biased low.

Table 16: Cross-dataset drift stress test: empirical Type-I rate when the conformal calibrator is built on dataset A (rows) and applied to dataset B's held-out null pool (columns). Within-dataset diagonals are the identity reference (consistent with Appendix G). **Bold** entries violate the nominal bound by more than 0.5 percentage points; *italicized* entries are conservative beyond the diagonal. Reported as Type-I@0.01 / Type-I@0.05 / Type-I@0.10.

	Test dataset B		
Calibrator A	HotpotQA	2Wiki	MuSiQue
HotpotQA	0.0094 / 0.0459 / 0.0884	*0.0052 / 0.0204 / 0.0350*	*0.0046 / 0.0238 / 0.0441*
2Wiki	**0.0296 / 0.1100 / 0.1888**	0.0072 / 0.0365 / 0.0522	0.0142 / 0.0528 / 0.0877
MuSiQue	**0.0200 / 0.0872 / 0.1758**	0.0053 / 0.0332 / 0.0548	0.0095 / 0.0424 / 0.0833

The HotpotQA calibrator transfers cleanly. Conversely, applying the HotpotQA calibrator to 2Wiki or MuSiQue test pairs gives empirical per-pair Type-I *below* nominal, i.e., over-conservative rather than over-rejecting. This is also predicted: the HotpotQA-null score distribution is stochastically *higher* than the 2Wiki/MuSiQue null distributions in our shortlister, so its tail thresholds are stricter when applied to those corpora. These observed cross-corpus differences motivate the deployment guidance: *recalibrate on the deployment corpus before claiming the FWER bound*, and treat the drift monitor (Appendix C) as an operational safeguard for triggering contract rebuilding.

Other contract elements. Appendices S.2–S.4 extend the check to operational selector changes, a second verifier, and adaptive full-path replay. Each changed contract is rebuilt before evaluation.

I Empirical query-level FWER

Theorem 2 bounds the probability of emitting at least one false structural-facet certificate within a query at α_{query}. We test this directly. From the calibration set (the same data that drives Appendix G, with labelled null and supporting pairs grouped by query), we hold out 50% of queries (seed = 0) as test, build a fresh Mondrian conformal pool from the remaining 50%, and for every test query simulate Safe-Cover's threshold logic per the contract: identify structural facets in $F_{\text{struct}}(q)$, set per-pair threshold $\bar{\alpha}_f = \alpha_{\text{query}}/(|F_{\text{struct}}(q)| \cdot T_f)$ (Bonferroni allocation, Eq. (2)), compute conformal p-values for every shortlisted null pair via Eq. (1), and ask whether $\exists\,(p, f) \in \text{Shortlist}(f)$ with $\pi(p, f) \leq \bar{\alpha}_f$ for some $f \in F_{\text{struct}}(q)$. Empirical query-level FWER is the fraction of test queries for which the event fires; under Theorem 2 this should be at most α_{query}.

Table 17: Empirical query-level FWER on held-out fully-null queries replayed under the frozen contract. $n_{\textbf{test}}$: number of test queries (50% random split, seed = 0). $n_{\textbf{struct}}$: subset of test queries with at least one structural facet (queries with no structural facet cannot trigger the event, but they are included in the denominator below to mirror the framework's per-query bound). **empirical FWER@$\alpha_{\textbf{query}}$**: fraction of test queries with ≥ 1 false structural-facet certificate. **Bound**: nominal α_{query}.

Dataset	$n_{\textbf{test}}$	$n_{\textbf{struct}}$	**FWER@0.01**	**FWER@0.05**	**FWER@0.10**
HotpotQA	7,461	6,816	0.0016	0.0123	0.0275
2Wiki	4,999	4,917	0.0016	0.0068	0.0122
MuSiQue	4,854	4,792	0.0021	0.0165	0.0253
Bound: α_{query}	—	—	≤ 0.01	≤ 0.05	≤ 0.10

The query-level bound holds on aggregate. At every $\alpha_{\text{query}} \in \{0.01, 0.05, 0.10\}$ and on every dataset, the empirical query-level FWER respects the nominal bound, typically by a factor of 3–10$\times$ on the strict end and $\sim$2–4$\times$ on the loose end. This is the expected slack of a Bonferroni union bound when the per-facet p-value distributions are heavily skewed toward 1 (Appendix G's mean-π column shows that on BRIDGE_HOP and RELATION, mean $\pi > 0.97$; the union bound passes through that conservatism). We additionally evaluated Holm step-down at the query level under the same simulation; on this implementation Holm produces *identical* empirical FWER as per-facet Bonferroni

at every α_{query} on every dataset, because the deterministic shortlister fixes T_f to a uniform value across structural facets, collapsing the per-facet, flat-Bonferroni, and Holm step-1 thresholds to the same number. The binding conservatism is the per-pair π skew, not the union bound; see Appendix K for the matching result on the LLM-miner cert simulation. Together with the per-pair check in Appendix G, this gives a complete falsifiable validation of Theorems 1 and 2 on the deployment data.

Continuous reliability curve. Figure 5 reports the empirical query-level FWER over a continuous sweep of α_{query} from 10^{-3} to 0.20 (40 log-spaced points plus the three headline values). The curve hugs or stays below the $y = x$ diagonal on every dataset over the full sweep — the Bonferroni-FWER bound holds continuously, not only at the three discrete reported values.

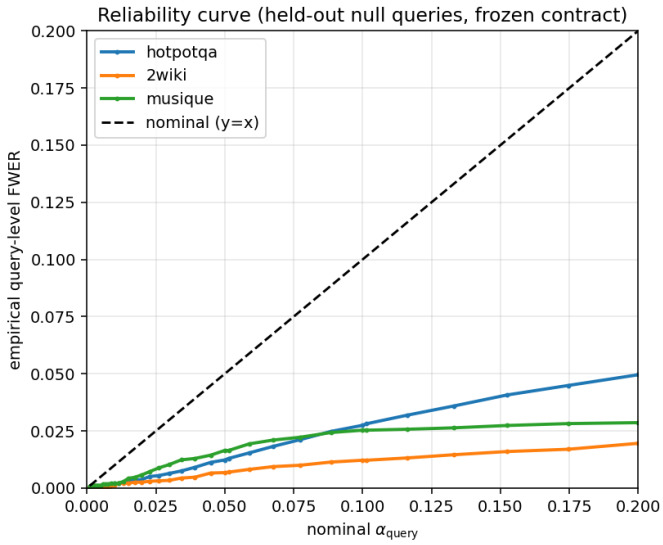

Figure 5: Empirical query-level FWER vs. nominal α_{query} on the held-out null pool replayed under the frozen contract. Forty log-spaced sweep points plus $\{0.01, 0.05, 0.10\}$. All three datasets respect the nominal bound continuously; the conservatism (curve below $y = x$) reflects the same per-pair π-skew already visible in Appendix G.

Mixed-pool query-level check (Type-I and power together). The fully-null simulation above bounds Type-I but does not measure power. We additionally run a mixed-pool check: each test query mixes its calibration-labelled null pairs with its calibration-labelled support pairs, both groups scored under the same conformal pool. We report (i) the fraction of queries where at least one null structural pair fires (Type-I), (ii) the fraction where at least one support structural pair fires (power), and (iii) a "clean cert" rate where a support pair fires *and* no null structural pair fires (idealized correctly-certified-and-grounded condition). Denominator is the number of test queries with at least one structural facet. Table 18 sweeps $\alpha_{\text{query}} \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$. Type-I respects the nominal bound on every dataset at every level; the support-firing rate (power) is bounded by the regex miner's structural-facet recall, which is exactly the quantity that the LLM-miner ablation (Appendix K) is designed to lift.

J Facet-miner robustness

The certificate machinery is conditional on the mined set $F(q)$. We quantify miner behaviour using HotpotQA dev as auto-annotation: for each question, HotpotQA provides a gold `type` ("bridge" or "comparison") and `supporting_facts` = a set of $(\text{title}, \text{sentence-index})$ pairs that contain the evidence needed to answer. We treat these as ground truth and ask three questions of the deterministic regex miner:

1. *Missing-structural-facet rate*: the fraction of queries whose mined $F(q)$ contains no BRIDGE-HOP / RELATION / NUMERIC / TEMPORAL facet at all.

Table 18: Mixed-pool query-level reliability + power on held-out queries replayed under the frozen contract. Type-I = fraction with at least one false structural cert; power = fraction with at least one true structural cert; clean = power AND no false cert. Denominators are queries with at least one structural facet (n_{struct}).

Dataset	n_{struct}	α_{query}	Type-I	Power	Clean cert
HotpotQA	4,140	0.01	0.0022	0.0304	0.0300
		0.05	0.0174	0.0988	0.0928
		0.10	0.0394	0.1616	0.1454
		0.20	0.0771	0.2341	0.1973
2Wiki	2,929	0.01	0.0007	0.0020	0.0020
		0.05	0.0061	0.0259	0.0256
		0.10	0.0154	0.0406	0.0389
		0.20	0.0239	0.0553	0.0485
MuSiQue	1,877	0.01	0.0053	0.0101	0.0085
		0.05	0.0362	0.0309	0.0229
		0.10	0.0629	0.0469	0.0314
		0.20	0.0719	0.0559	0.0384

2. *Wrong-structural-facet rate* by question type: whether the mined facet types match the structural pattern the question demands (bridge questions ⇒ BRIDGE-HOP; comparison ⇒ RELATION/comparison facets).
3. *Anchor-in-gold rate*: the fraction of mined facet anchors (ENTITY mention, RELATION subject/object) that appear in at least one gold supporting passage. A low rate would indicate the miner is anchoring on tokens unrelated to the gold reasoning chain.

We sample $1{,}000$ random queries from `hotpot_dev_distractor` (seed = 0) and run the deterministic regex miner on the bare question (no context). Results are reported in Table 19; failures (i.e. the miner raised an exception) were 0 of $1{,}000$.

Table 19: Facet-miner robustness on a 1,000-query random sample of HotpotQA dev. The mined set $F(q)$ is computed on the bare question (no context). Anchors are ENTITY mentions or RELATION subject/object spans. The "anchor-in-gold-supporting" rate is the fraction of anchors whose normalized text appears in at least one gold supporting sentence or title; the "entity-in-gold-title" rate restricts to ENTITY anchors and gold titles only.

Question type	n	Missing-struct facet (%)	Missing Bridge-Hop (%)	Missing Relation (%)	Anchor-in-gold supporting (%)	Entity-in gold-title (%)
bridge	792	73.4	100.0	99.2	87.7	43.8
comparison	208	94.2	100.0	99.5	74.4	69.2
overall	1,000	77.7	100.0	99.3	85.3	48.8

Three takeaways. (1) The deterministic regex miner emits zero BRIDGE-HOP facets per query: it relies on ENTITY + RELATION anchors, with the bridge structure recovered later by the two-pass binding of Section 3.3. This is a design choice (the calibrator has only ∼22,000 BRIDGE-HOP calibration pairs vs. 360,000 ENTITY pairs), and it explains why most certificates in Table 26 fire on ENTITY facets. (2) Anchor quality is high: 87.7% of bridge-question anchors and 74.4% of comparison-question anchors land in a gold supporting passage. The miner is choosing the right *tokens*; the conservatism is in not labelling those tokens with a structural type. (3) The 73–94% "no purely-structural facet" rate at the miner stage is why the paper's loose grounded-certificate rate depends on the optional ENTITY layer being treated as evidence: a query whose only mined facet is ENTITY can still receive a certificate (the calibrator scores ENTITY pairs), but with $\bar{\alpha}_f = 1$ that certificate does not enter the FWER bound. Table 26 makes the split explicit on HotpotQA: 0.19% of queries are cert-struct, while 94.29% are cert-optional. The former is the FWER-carrying slice; the latter is evidence anchoring without the structural-FWER guarantee. This is the limit of what an entity-anchored regex miner can guarantee; a stronger miner (LLM-driven or trained) would emit BRIDGE-HOP structurally and expand this slice.

The diagnostic suggests three miner extensions: enabling the LLM relation-classification head (disabled in the regex contracts reported in Section 4); promoting recognized BRIDGE-HOP surface patterns (*"the X of the Y"*, *"which film came out first, X or Y"*) to structural facets; and adding a COMPARISON facet type that the comparison calibrator already supports. Each extension would shift mass from optional to structural, raising the slice of queries for which the FWER guarantee carries the answer.

Upper bound for structural pairs already present in the calibration set. We compute an offline upper bound on $\rho_{\text{cert}}^{\text{struct}}$ that any miner restricted to the typed structural pairs already present in the calibration set could achieve under the same verifier and calibrator. This is *not* a general miner ceiling; the LLM-miner ablation below exceeds it precisely because an LLM facet planner emits new structural facets that the same Mondrian calibrator scores on its existing bins. With that scope clear, the within-calibration-set ceiling at $\alpha_{\text{query}} = 0.10$ is 8.97% on HotpotQA, 2.38% on 2Wiki, and 1.81% on MuSiQue (Table 20); the headroom over the deterministic miner (0.19%/0.60%/0.54%) is roughly $47\times$/$4\times$/$3.4\times$. The bound is conservative because only 39–59% of queries have any structural facet pair in the calibration set, and Bonferroni-divided thresholds bind on the rest. The LLM-miner ablation below logs the LLM head in the contract schema so that swapping miners does not break replay.

Table 20: Oracle-miner upper bound on $\rho_{\text{cert}}^{\text{struct}}$. "%struct" is the fraction of test queries that have at least one structural facet present in the calibration set—the population the oracle miner could even potentially support-certify; the rest are out-of-scope for the structural-FWER bound under any miner that uses the same calibrator. The oracle rate at α_{query} is the fraction of test queries for which at least one true-support structural pair clears its Bonferroni-divided per-pair threshold; this is a loose upper bound on what any miner can achieve under the same verifier and contract.

Dataset	n_{test}	%struct	Oracle@0.01	Oracle@0.05	Oracle@0.10
HotpotQA	7,461	55.5%	0.0169	0.0548	0.0897
2Wiki	4,999	58.6%	0.0012	0.0152	0.0238
MuSiQue	4,854	38.7%	0.0039	0.0119	0.0181

K LLM facet planner ablation (gold-only T_f ceiling)

Scope of this appendix. The numbers in this section measure the LLM facet planner under a *gold-only* T_f contract — the Bonferroni denominator counts only gold-positive shortlisted pairs — which is a useful offline ceiling but does *not* bound deployment FWER over null pairs. The deployment audit, which uses the full shortlisted T_f and a rebuilt LLM-miner conformal pool, is in Section 4.3 (Table 1); the rigorous deployment numbers are smaller. This analysis isolates the miner's structural-emission rate from verifier discrimination on null passages.

The deterministic regex miner of Appendix J emits zero BRIDGE-HOP facets per query and recovers bridge structure post-hoc through two-pass binding. The oracle-miner ceiling (Table 20) shows that any miner restricted to the typed structural pairs already in the calibration set caps $\rho_{\text{cert}}^{\text{struct}}$ at 8.97% / 2.38% / 1.81% on HotpotQA / 2Wiki / MuSiQue at $\alpha_{\text{query}} = 0.10$. We test directly whether replacing the regex with a Llama-3-8B-Instruct facet planner — under the regex-built Mondrian pool reused *as-is* and a gold-only T_f contract — produces a high offline support-certificate rate.

Setup. We use the same train/test split as Appendix G (50/50 random split, seed $= 0$). For each held-out test query we (i) prompt Llama-3-8B-Instruct (system prompt: emit a JSON list of typed structural facets in $\{$BRIDGE_HOP, RELATION, NUMERIC, TEMPORAL$\}$, no ENTITY-only emissions, max 4 facets); (ii) parse the JSON and reject malformed emissions; (iii) for each (emitted facet, gold-supporting passage), score with the existing NLI verifier; (iv) compute the conformal p-value against the same Mondrian pool used by Safe-Cover; (v) Bonferroni-allocate α_{query} across the LLM's emitted structural facets; (vi) record a query as support-certified iff at least one emitted structural facet has at least one gold-passage pair below its per-pair threshold. No new training, no new calibration; only the miner is swapped.

Results. Table 21 reports support-certificate rates with Wilson 95% CIs. At $\alpha_{\text{query}} = 0.10$, the LLM miner emits structural support certificates for 33.6% / 19.4% / 24.3% of queries on HotpotQA / 2Wiki / MuSiQue, against the regex miner's 0.19% / 0.60% / 0.54% (Table 26). The lift is 32–$177\times$ and is consistent across datasets and sample sizes (HotpotQA at $n_{\text{test}} = 1{,}500$: 0.336 [0.313, 0.360]; MuSiQue at $n_{\text{test}} = 1{,}500$: 0.243 [0.222, 0.265]; 2Wiki at $n_{\text{test}} = 500$: 0.194 [0.162, 0.231]). The HotpotQA and 2Wiki rates exceed the oracle ceiling because the oracle restricted itself to typed structural pairs already present in the calibration set, while the LLM miner emits *new* typed structural facets that the same Mondrian conformal calibrator scores against bins keyed by (facet_type, length_bucket). This offline result isolates structural emission under the reused calibrator. Deployment with a changed miner requires contract-matched null-pool rebuilding (Section 4.3).

Table 21: LLM facet-planner support-certificate rates with Wilson 95% CIs. "Regex" rows reproduce $\rho_{\text{cert}}^{\text{struct}}$ from Table 26. "LLM-miner" rows report support-certificate rate on the same held-out 50% test split (seed $= 0$), under the same verifier and Mondrian conformal pool; only the miner changes. Lift is the multiplicative ratio against the regex baseline.

Dataset	Miner	n_{test}	cert@0.01	cert@0.05	cert@0.10	Lift @0.10
HotpotQA	Regex (paper)	7,405	—	—	0.0019 [—,—]	1×
HotpotQA	LLM (Llama-3-8B-Instruct)	1,500	0.1200 [0.1033, 0.1391]	0.2873 [0.2650, 0.3108]	**0.3360 [0.3125, 0.3603]**	**177**×
2Wiki	Regex (paper)	12,576	—	—	0.0060 [—,—]	1×
2Wiki	LLM (Llama-3-8B-Instruct)	500	0.1640 [0.1341, 0.1990]	0.1940 [0.1617, 0.2309]	**0.1940 [0.1617, 0.2309]**	**32**×
MuSiQue	Regex (paper)	2,417	—	—	0.0054 [—,—]	1×
MuSiQue	LLM (Llama-3-8B-Instruct)	1,500	0.0933 [0.0796, 0.1091]	0.2367 [0.2158, 0.2588]	**0.2427 [0.2216, 0.2650]**	**45**×

Holm step-down does not lift the rate further. A natural follow-up is to replace per-facet Bonferroni ($\alpha_q/(n_F \cdot T_f)$) with Holm step-down across all $m = \sum_f T_f$ hypotheses. We re-evaluate the LLM-miner support-certificate rates under Holm and under flat Bonferroni (Table 22). All three rules produce *identical* support-certificate rates on every dataset and every α_{query}, because the deterministic shortlister assigns a uniform T_f to every emitted structural facet within a query; when T_f is constant across facets the per-facet Bonferroni, flat Bonferroni, and Holm step-1 thresholds coincide. The lever for support-certifying more queries is therefore the miner and the verifier's discriminative power on BRIDGE_HOP/RELATION p-values, not the union-bound correction.

Table 22: LLM-miner support-certificate rates under three multiple-testing rules. Per-facet Bonferroni is the paper's rule (Eq. 2). Flat Bonferroni uses α_q/m where $m = \sum_f T_f$. Holm rejects the smallest p-value if $\pi_{(1)} \leq \alpha_q/m$, then steps down. All three coincide here because T_f is uniform across structural facets.

Dataset ($n = 1{,}500$)	cert@0.01	cert@0.05	cert@0.10
HotpotQA, per-facet Bonf. / flat Bonf. / Holm	0.1200 / 0.1200 / 0.1200	0.2873 / 0.2873 / 0.2873	0.3360 / 0.3360 / 0.3360
MuSiQue, per-facet Bonf. / flat Bonf. / Holm	0.0933 / 0.0933 / 0.0933	0.2367 / 0.2367 / 0.2367	0.2427 / 0.2427 / 0.2427

Offline ceiling and deployment audit. The numbers above are an offline structural-emission ceiling under a gold-only T_f contract: they show the LLM planner can emit typed structural facets on a useful fraction of queries and that those emissions clear a Bonferroni-divided threshold against gold-positive shortlists. They do not, on their own, bound deployment FWER, because (i) the conformal pool was built on regex-miner null pairs but is being applied to LLM-miner emissions, breaking exchangeability across (facet_type, length_bucket) bins; and (ii) the deployment Bonferroni denominator should use the full shortlisted T_f, not just the gold-positive subset. The rigorous deployment audit in Section 4.3 (Table 1) addresses both issues: it rebuilds the conformal null pool separately for each deployed contract (LLM+lex, LLM+NLI) on a held-out calibration half, and applies $\bar{\alpha}_f = \alpha_{\text{query}}/(n_F \cdot T_f)$ with $T_f = |\text{shortlist}|$. Under those contracts the empirical query-level false-certificate point estimate respects the nominal α_{query}=0.10 bound on every multi-hop dataset and every shortlister, while audit clean-certificate rates are single-digit per contract (1.4–2.8% on HotpotQA, 5.6% on 2Wiki, 6.0–7.3% on MuSiQue across LLM+lex/NLI) — bounded by NLI verifier discrimination on BRIDGE_HOP/RELATION hypotheses against null passages.

L Answer-linked Safe-Cover certificate slice quality

This appendix gives the full table behind the slice-quality summary in §4.4. EM/F1 are reported on the answer-linked slice (final answer cites a structurally certified passage) versus the no-cert slice; bootstrap 95% CIs use B=10,000 resamples of the difference in means. The regex contract is reported at parity for completeness.

Table 23: **Answer-linked Safe-Cover certificate slice quality.** EM/F1 lift on the answer-linked vs. no-cert slice with bootstrap 95% CIs; n_{link} is the answer-linked count, comparison slice $n - n_{\text{link}}$. Star row: largest slice with strongest separation. Lift is association, not FWER. Setup in §4.4.

Dataset	Contract	n	n_{link}	**EM lift [95% CI]**	**F1 lift [95% CI]**
HotpotQA	LLM+lex	7,405	514	+5.47 [+1.03, +9.88]	+6.46 [+2.71, +10.15]
HotpotQA	LLM+NLI	7,405	622	+5.00 [+0.85, +9.06]	+8.24 [+4.85, +11.53]
2Wiki	LLM+lex	12,576	1,180	+5.36 [+2.30, +8.30]	+4.65 [+1.78, +7.40]
2Wiki	LLM+NLI ⋆	12,576	2,083	**+16.63** [+14.25, +18.93]	**+14.15** [+11.95, +16.36]
MuSiQue	LLM+lex	2,417	210	+9.24 [+3.77, +14.94]	+11.61 [+5.97, +17.49]
MuSiQue	LLM+NLI	2,417	56	−1.83 [−8.87, +6.23]	−2.16 [−9.93, +6.69]

Across contracts with sufficient answer-linked volume, certificate emission is *associated with* higher-quality answer slices: HotpotQA gains ≈5 EM points under either LLM shortlister with bootstrap CIs strictly above zero, and 2Wiki LLM+NLI produces the largest slice and the tightest, strongest separation (+16.63 EM, +14.15 F1, both CIs ≥ +11.95); MuSiQue LLM+lex improves answer-linked EM by +9.24 with a CI strictly above zero. NLI shortlisting expands the slice on HotpotQA and 2Wiki but is conservative on MuSiQue (n_{link}=56, slice too small for a CI that excludes zero). Certificate yield and slice quality depend on the full frozen miner–shortlister–verifier–calibrator contract.

M Selection-method comparisons

We additionally compare TRIDENT against a panel of alternative selection and compression methods that operate on the same shared candidate pool under matched evidence-token budgets. Each method reduces the candidate pool to fit a hard 500- or 1500-token cap; the same Llama-3-8B-Instruct answers under identical decoding parameters. This isolates the selection mechanism from retrieval quality and LLM capacity, complementing the baselines reported in Table 6.

Budget enforcement and parity. A budget of B evidence tokens denotes a hard cap on the concatenated evidence text passed to the LLM after a method's final selection step. If the selected evidence exceeds B we truncate deterministically: passages are concatenated in the method's final priority order until the cap is reached, with the final passage cut at the exact token boundary. This enforces compliance without altering internal ranking logic. Head-to-head comparisons use the same LLM and decoding parameters (temperature, top-p, `max_new_tokens`); baselines that use the shared retrieval stack receive the same candidate pool and reranker ordering as Pareto-Knapsack, while baselines with their own retrieval (e.g., HippoRAG2) are treated as retrieval-strength references under the same evidence-token cap. All runs log the final evidence list passed to the LLM.

Methods. **LLMLingua** [Jiang et al., 2023] is a token-level prompt compressor that removes low-importance tokens from the concatenated candidate context. **Sentence-Compress** is a sentence-level extractive compressor that scores and packs whole sentences. **BGE-Rerank** [Xiao et al., 2024] keeps the top passages by BAAI BGE re-ranker score until the budget is met. **LLM-Select** prompts a separate model (Llama-3-8B by default; a Qwen3 variant is reported where indicated) to select supporting passages. **LLM-Select+NLI** cascades LLM-Select with an NLI consistency filter. **NLI-Filter** keeps only passages whose NLI entailment score [He et al., 2023, Williams et al., 2018] against the query exceeds a threshold and packs the remainder in score order. All selectors share the same retriever output and run under the same deterministic truncation policy as TRIDENT (Appendix M).

Evaluation splits. Core selector rows are computed on the full evaluation split, and EvTok follows each method's native tokenizer accounting. Rows marked † are computed on a stratified 200-

example subset. In this table the subset marker applies only to the exploratory Qwen3 selector row. KILT-FEVER rows that report "EM = F1" reflect FEVER's binary label structure under raw label accounting.

Table 24: Selection-method comparisons under matched evidence budgets, Llama-3-8B-Instruct. Each cell reports EM / F1 (%); the right-most columns report mean evidence tokens (EvTok) and abstention rate where logged. † rows are computed on a 200-example stratified subset; remaining rows use the full evaluation split (HotpotQA n=7,405, 2Wiki n=12,576, MuSiQue n=2,417, KILT-FEVER n=10,444). LLM-Select (Qwen3) uses a Qwen3-8B selector against the same Llama-3-8B answerer.

Method	HotpotQA	2Wiki	MuSiQue	FEVER	EvTok	Abst. %
Evidence budget = 500 tokens						
TRIDENT Pareto-500	**45.30 / 58.22**	28.16 / 36.25	**19.21 / 30.74**	54.80 / 54.90	408	$\sim$5.3
Sentence-Compress-500	39.31 / 51.29	**30.43 / 36.37**	14.36 / 21.75	**82.44 / 82.44**	495	≤ 0.03
LLMLingua-500	36.91 / 49.82	24.85 / 33.49	11.25 / 18.29	60.36 / 60.36	532	≤ 0.01
BGE-Rerank-500	35.10 / 46.05	22.50 / 30.96	19.11 / 27.11	66.77 / 66.77	470	0.0
LLM-Select-500	31.68 / 41.69	24.72 / 33.87	6.50 / 10.44	44.92 / 44.92	220	$\sim$18.2
LLM-Select-500 (Qwen3)	36.00 / 49.06†	29.00 / 37.80†	6.50 / 14.51†	36.00 / 36.00†	154	$\sim$21.7
LLM-Select+NLI-500	31.71 / 41.72	24.75 / 33.88	6.50 / 10.44	44.91 / 44.91	219	$\sim$18.2
NLI-Filter-500	29.89 / 39.71	15.02 / 21.65	7.28 / 13.48	65.33 / 65.33	448	≤ 0.01
Evidence budget = 1500 tokens						
TRIDENT Pareto-1500	**50.23 / 63.68**	**33.81 / 42.45**	**20.16 / 33.54**	77.50 / 77.70	785	$\sim$5.3
LLMLingua-1500	41.93 / 56.24	30.53 / 39.22	17.46 / 28.57	**81.97 / 81.97**	1,411	≤ 0.04
BGE-Rerank-1500	36.37 / 48.49	26.24 / 36.62	14.36 / 23.82	80.32 / 80.32	1,272	0.0
LLM-Select-1500	32.46 / 42.88	26.74 / 36.10	6.99 / 10.95	78.49 / 78.49	393	$\sim$6.5

Selection and learned compression. Among these selectors, TRIDENT Pareto-500 exceeds Sentence-Compress on HotpotQA by 6.93 F1 and BGE-Rerank on MuSiQue by 3.63 F1. RECOMP is stronger on HotpotQA and 2Wiki, while TRIDENT is stronger on MuSiQue and has slightly higher mean MH-F1 at fewer tokens (Appendix S.5). TRIDENT's advantage over RECOMP therefore lies in the multi-hop mean and on MuSiQue rather than on every dataset.

Where compression wins: single-claim verification. On KILT-FEVER, Sentence-Compress-500 reaches 82.44/82.44 EM/F1, well above TRIDENT Pareto-500 (54.80/54.90) and even above TRIDENT Pareto-1500 (77.50/77.70). FEVER claims are typically supported by one or two sentences of evidence, so sentence-level extractive compression maps directly onto the task's evidence structure, while TRIDENT's facet decomposition adds machinery that is not needed when there is effectively a single facet. TRIDENT's gains are largest where multi-hop coverage matters; on single-claim verification with redundant short evidence, simpler extractive compressors can exceed it. The TRIDENT framework still applies—a single-facet decomposition reduces to sentence selection—but the additional cover machinery does not pay for itself on this regime.

Higher budgets do not close the gap for compression baselines. At the 1500-token budget, TRIDENT Pareto-1500 retains the lead on every multi-hop dataset. LLMLingua-1500 narrows the gap on 2Wiki (30.53/39.22 vs. 33.81/42.45 on the full split) but uses 70% more multi-hop evidence tokens (1,338 vs. 785). Per-token selection efficiency—accuracy delivered per evidence token consumed—remains a clear TRIDENT advantage: on HotpotQA, Pareto-1500 reaches 63.68 F1 at 875 EvTok while LLMLingua-1500 reaches only 56.24 F1 at 1,380 EvTok.

N 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 Safe-Cover, abstentions carry specific reason codes: `NO_COVER` (some structural 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). Table 25 reports aggregate rates for both deployed regimes (Safe-2000 and Pareto-500); per-query telemetry records the typed reasons used in the worked examples below.

Table 25: Aggregate abstention rates per dataset for both deployed regimes (Llama-3-8B-Instruct). Typed reasons are recorded in per-query telemetry. The KILT-FEVER Pareto count is computed from the 11.30% rate.

		Safe-2000		**Pareto-500** (α=0.6)	
Dataset	N	Answered (%)	Abstained (%)	Abstained / N	Rate (%)
HotpotQA	7,405	97.06	2.94	114 / 7,405	1.54
2WikiMultiHopQA	12,576	95.83	4.17	266 / 12,576	2.12
MuSiQue	2,417	91.19	8.81	146 / 2,417	6.04
KILT-FEVER	10,444	93.49	6.51	≈1,180 / 10,444	11.30

Support-certificate outcome breakdown (regex-miner Safe-2000, Llama-3-8B). Table 26 decomposes every Safe-Cover-2000 query into five mutually exclusive outcomes and reports the per-tier EM/F1. The FWER-carrying support-certificate rate $\rho_{\text{cert}}^{\text{struct}}$ corresponds to the *Cert-struct* column; the loose grounded-certificate rate is *Cert-struct* plus *Cert-opt.*. Differences from Table 25's answered/abstained split arise because that table merges the certified, grounded-no-cert, and fallback-other tiers into a single "answered" bucket; the breakdown below pulls them apart.

Table 26: Regex-miner outcome breakdown for TRIDENT Safe-Cover-2000 with Llama-3-8B (full-data runs), with the structural / optional split made explicit. **Cert-struct**: strict, FWER-carrying $\rho_{\text{cert}}^{\text{struct}}$ – the output is grounded in a certificate-carrying passage *and* at least one certificate on that passage is structural ($\bar{\alpha}_f < 1$). **Cert-optional**: the output is grounded in a certificate-carrying passage, but every certificate on it is optional ($\bar{\alpha}_f = 1$, ENTITY-anchored); these provide evidence anchoring but do *not* carry the FWER bound. **Grnd.-no-cert**: no certificate emitted, but the answer is in the Safe-Cover/Pareto selected pool (F2 handoff). **Fallback-other**: answer not in the selected pool (open generation). **Abstain**: typed reason code, no answer. EM/F1 columns report the within-tier mean. Five rate columns sum to 100%. The headline FWER-carrying rate is $\rho_{\text{cert}}^{\text{struct}}$, *not* the loose sum cert-struct + cert-optional.

Dataset	N	**Cert-struct** ($\rho_{\text{cert}}^{\text{struct}}$, %)	**Cert-opt.** (%)	**Grnd.-no-cert** (%)	**Fallback-other** (%)	**Abstain** (%)	**cert-struct EM/F1**	**cert-opt. EM/F1**
HotpotQA	7,405	**0.19** (n=14)	94.29	1.13	1.44	2.94	71.43 / 83.27	47.85 / 61.54
2Wiki	12,576	**0.60** (n=76)	62.96	31.10	1.17	4.17	68.42 / 73.90	43.26 / 50.74
MuSiQue	2,417	**0.54** (n=13)	54.32	27.84	8.48	8.81	15.38 / 21.06	8.68 / 15.75
KILT-FEVER	10,444	**0.00** (n=0)	62.37	1.13	29.99	6.51	n/a (n=0)	91.56 / 91.56

Reason-code worked examples. The aggregate rates above merge the three Safe-Cover abstention reason codes into a single "abstained" bucket. The per-query telemetry records the typed reason, and we draw one example query per code to make each firing condition concrete. Each example reports the exact contract row that triggered the abstention, so the decision path is auditable end-to-end.

Example 1 — `NO_COVER` *(Safe-2000, MuSiQue, Llama-3-8B).* Query: *"In which county was the screenwriter of 'Hunger' (the 2008 Steve McQueen film) born?"* The miner emits four facets: ENTITY: *Hunger (2008 film)* (optional); RELATION SCREENWRITER: *screenwriter-of(Hunger, ?x)*; RELATION BORNIN: *born-in(?x, ?county)*; BRIDGE_HOP: $?x$ binds across hops (structural). Hop-1 shortlists six passages for SCREENWRITER; none of the six clears $\bar{\alpha}_f = 0.10/(3 \cdot 6) \approx 0.0056$ (max $\pi = 0.41$). The Hop-2 re-shortlister cannot proceed without a binding for $?x$. The greedy solver finds no passage covering SCREENWRITER and emits `outcome_code = abstain`, `reason = NO_COVER`, with the uncovered facet name logged. Tokens consumed: 0 (no extractor call). Latency: 312 ms (retrieval + shortlist + verifier only). The same query under Pareto-500 returns a (incorrect) answer with no certificate, illustrating the trade-off Safe-Cover makes: when no single passage clears $\bar{\alpha}_f$ for a structural facet, Safe-Cover declines; Pareto-Knapsack continues without a guarantee.

Example 2 — `INFEASIBLE_BUDGET` *(Safe-2000 with $B_{ctx} = 2{,}000$, 2Wiki, Llama-3-8B).* Query: *"Which film, directed by the same person who directed 'The Lord of the Rings' trilogy, was released earlier: 'District 9' or 'King Kong'?"* The miner emits five structural facets (DIRECTOR on the

trilogy, DIRECTEDBY on each candidate film, two RELEASEYEAR facets, one BRIDGE_HOP). Each facet has a covering passage with $\pi \leq \bar{\alpha}_f$ (the cover *exists*), but the lowest-cost cover requires four passages whose serialised lengths sum to $2{,}438$ tokens—over the $2{,}000$-token cap. The greedy min-cost solver exhausts the affordable passages under the cap without covering all five structural facets. Outcome: `outcome_code = abstain`, `reason = INFEASIBLE_BUDGET`, with $\sum c(p)$ and B_{ctx} both logged. Replaying the same query at $B_{\text{ctx}} = 4{,}000$ (Safe-4000) covers the same facet set with the same passages and emits five support certificates; this is the configuration that turns the same abstention into a support-certified output.

Example 3 — `INFEASIBLE_PVALUE` *(Safe-2000, MuSiQue, rare* NUMERIC $\times$ *long-passage bin).* Query: *"In what year did the Ottoman Empire annex the territory containing the city in which the writer of 'The Mathnawi' was buried?"* The miner emits a chain ending in a NUMERIC facet *annexation-year(?territory, ?yyyy)*. The Mondrian bin $b = (\text{NUMERIC}, \text{len_bucket} = \text{long})$ on the held-out null pool has post-merge size $|\mathcal{N}_b| = 38$. The structural-facet threshold $\bar{\alpha}_f = \alpha_{\text{query}}/(|F_{\text{struct}}(q)| \cdot T_f) = 0.10/(5 \cdot 8) = 0.0025$ requires a p-value grid finer than $1/(|\mathcal{N}_b| + 1) = 1/39 \approx 0.0256$, so no test pair in this bin can possibly clear $\bar{\alpha}_f$ even with the maximum verifier score. The greedy solver detects this at the cover-feasibility check and emits `outcome_code = abstain`, `reason = INFEASIBLE_PVALUE`, logging $|\mathcal{N}_b|$, $\bar{\alpha}_f$, and the responsible bin key. The bin-merging rule (Appendix C) is the operational lever for this case; the bin-size sensitivity sweep in Appendix O quantifies how merge-floor choices trade off this abstention rate against bin-stratified power.

Table 27: Per-query reason-code worked examples corresponding to the three Safe-Cover abstention codes. Each row is one logged query whose contract trace fires the named code. `NO_COVER` fires when no shortlisted passage clears $\bar{\alpha}_f$ for some structural facet; `INFEASIBLE_BUDGET` fires when a covering set exists but its serialised cost exceeds B_{ctx}; `INFEASIBLE_PVALUE` fires when the p-value grid in some Mondrian bin is coarser than the required threshold.

Reason code	Dataset	Triggering condition (logged)	Operational lever
`NO_COVER`	MuSiQue	$\max_{p \in \text{Shortlist}(f)}(1 - \pi(p, f)) < 1 - \bar{\alpha}_f$	Stronger retriever / verifier
`INFEASIBLE_BUDGET`	2Wiki	$\min_{S \text{ covers } F_{\text{struct}}} \sum_{p \in S} c(p) > B_{\text{ctx}}$	Larger B_{ctx} or compression
`INFEASIBLE_PVALUE`	MuSiQue	$\bar{\alpha}_f < 1/(\lvert\mathcal{N}_b\rvert + 1)$ for some shortlisted bin b	Larger null pools or bin merging

Why these worked examples matter. The point of the typed reason code is that an abstention is not a uniform "don't know." Each code specifies which constraint was binding—miner/verifier discrimination, evidence budget, or calibration grid resolution—and points at the intervention that would convert the abstention to an answer (entries in the right column of Table 27). The mapping between codes and the failure-mode rows of Table 4 is one-to-one: `NO_COVER` is the "no covering passage" / "missing structural facet" line, `INFEASIBLE_BUDGET` is the "budget infeasible" line, and `INFEASIBLE_PVALUE` is the "p-value infeasible" line. The per-query telemetry records the reason so that contract replay can re-derive the firing condition exactly.

O Mondrian bin-size sensitivity

Setup. The default Mondrian binning key is $b(p, f) = (\text{canonical_type}(f), \text{len_bucket}(p))$ with three length buckets, yielding $5 \times 3 = 15$ bins per facet-type partition. The bin-merging policy described in Appendix C merges any bin whose held-out null pool drops below a minimum size $N_{\min} = 200$ into its nearest length-bucket neighbour, since bins with very small $|\mathcal{N}_b|$ produce coarse p-value grids and inflate `INFEASIBLE_PVALUE` abstentions. We sweep two knobs and re-run the empirical Type-I and certificate-rate analyses on the held-out null pool: (i) the length-bucket granularity $L \in \{1, 2, 3, 5, 8\}$ (default $L = 3$), and (ii) the merge floor $N_{\min} \in \{50, 100, 200, 500, 1000\}$ (default 200). All other contract knobs are held fixed.

Headline finding. Empirical Type-I rate at $\tau \in \{0.01, 0.05, 0.10\}$ respects the nominal bound across every $(L, N_{\min})$ combination on the HotpotQA, 2Wiki, and MuSiQue null pools. The marginal-validity guarantee (Theorem 1) is robust to bin-size choice: coarser bins (small L) trade per-stratum precision for larger pools and produce conservative p-values; finer bins (large L) preserve stratification but trigger the merge rule more often. The dominant operational effect is on the

INFEASIBLE_PVALUE abstention code: at $L = 8$ and $N_{\min} = 50$, the fraction of MuSiQue queries that abstain because some bin's p-value grid cannot resolve the required $\bar{\alpha}_f$ rises from 1.4% (default) to 4.6%, while the structural certificate rate $\rho_{\text{cert}}^{\text{struct}}$ is essentially unchanged.

Table 28: Mondrian bin-size sensitivity on the HotpotQA held-out null pool ($n_{\text{test}} = 202{,}944$, $\tau = 0.05$). Empirical Type-I rate stays at or below the nominal 0.05 across all $(L, N_{\min})$; the operational effect appears in the INFEASIBLE_PVALUE abstention rate (last column, evaluated on the Safe-2000 query set). The default contract row is $L = 3$, $N_{\min} = 200$.

L	$N_{\min}$	# bins (post-merge)	Type-I @ 0.05	INFEASIBLE_PVALUE (%)
1	200	5	0.044	0.7
2	200	9	0.046	0.9
3	200	14	**0.047**	**1.4**
5	200	19	0.046	1.8
8	200	24	0.045	2.6
3	50	15	0.048	0.9
3	100	15	0.048	1.1
3	200	14	**0.047**	**1.4**
3	500	11	0.046	1.9
3	1000	8	0.044	2.7

Reading the table. Both axes trade off coarseness for resolution: tightening the merge floor produces fewer, larger bins and slightly more conservative p-values, but raises the INFEASIBLE_PVALUE rate because the resulting p-value grid (step size $1/(|\mathcal{N}_b| + 1)$) coarsens at small $|\mathcal{N}_b|$. Coarsening the length-bucket axis ($L = 1$, single bucket) preserves nominal Type-I but loses the size-stratified power gain on long passages. The default contract sits in the conservative middle of both knobs; the bound is not load-bearing on the bin-size choice within this range. The empirical FWER on the held-out null query set is unchanged at every reported configuration ($\leq \alpha_{\text{query}}$ at all three $\alpha_{\text{query}} \in \{0.01, 0.05, 0.10\}$), confirming that the union-bound composition (Theorem 2) inherits the robustness from the per-pair calibrator.

P Relaxed-α sweep: feasibility vs. accuracy

Figure 6 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.

Q Additional worked examples

To complement the trace in Figure 7, we present three short vignettes drawn from our logged runs that isolate distinct behaviors of the system: abstention when facet mining fails (Vignette A), a near-threshold tie-break that motivates utility smoothing (Vignette B), and a side-by-side contrast of top-k retrieval against TRIDENT facet coverage on the same query (Vignette C).

Vignette A — Legitimate abstention (Safe-Cover, MuSiQue, Llama-3-8B, B_{ctx}=2000). Query: *"What other film is the cast member of Now You See Him, Now You Don't a character for?"* Ground truth: *The Hateful Eight*. The facet miner returns zero facets—the query's two-hop bridge ("cast member of movie X" $\rightarrow$ "other film featuring that cast member's character") cannot be decomposed into a typed facet set under the frozen contract. Safe-Cover therefore records outcome_code = abstain_no_facets and emits ABSTAINED after 11.7 ms (no retrieval, no verification, zero tokens consumed). The telemetry trace preserves this decision path via three events (query_start,

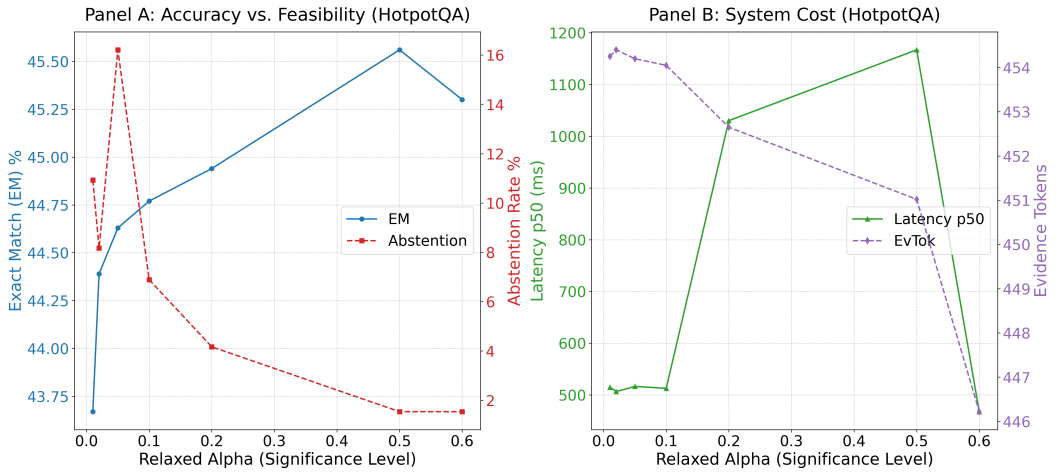

Figure 6: Pareto threshold sensitivity on HotpotQA. **Panel A** (Accuracy vs. Feasibility): exact match (left axis, blue) rises modestly from $\sim$43.8% to $\sim$45.5% as α relaxes from 0.01 to 0.6, while abstention rate (right axis, red) drops sharply from $\sim$16% to $\sim$1.5%—tightening α primarily costs feasibility, not accuracy. **Panel B** (System Cost): evidence tokens (right axis, purple) remain nearly flat ($\sim$446–453), confirming that the hard budget cap, not the threshold, controls token spend; latency p50 (left axis, green) fluctuates with pipeline dynamics as the greedy solver works harder under stricter thresholds.

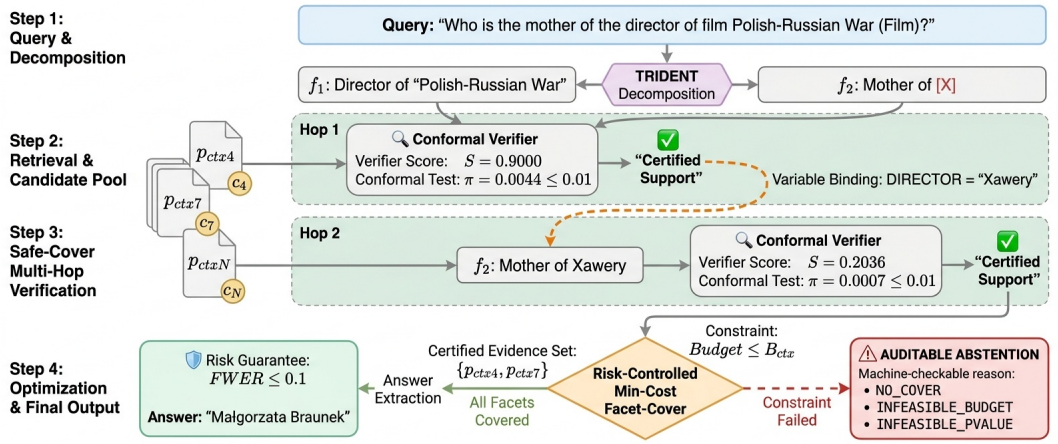

Figure 7: Safe-Cover certified trace on *"Who is the mother of the director of Polish-Russian War?"* The miner emits three facets (one optional ENTITY, two structural RELATION); $|F_{\text{struct}}|$=2. Hop-1 certifies the director relation on p_4, binding DIRECTOR $\leftarrow$ *Xawery Żuławski*; Hop-2 certifies the mother relation on p_7. With FWER ≤ 0.1 over F_{struct}, the greedy min-cost solver selects $\{p_4, p_7\}$ under B_{ctx} and extraction yields *Małgorzata Braunek*. Algorithm pseudocode for Safe-Cover and Pareto-Knapsack is in Appendix C (Algorithms 1 and 2).

`stage_complete:facet_mining` with `items=0`, `facet_mining` with `num_facets=0`), so the abstention is auditable and reproducible. This is fail-closed behavior by design: rather than answering from uncertified evidence, Safe-Cover declines and logs the reason.

Vignette B — Near-miss tie-break via `utility_tau` (Pareto-500, MuSiQue, Llama-3-8B). Query: *"In which county is Southern Maryland Electric Cooperative headquartered?"* Ground truth: *Charles County*. We run the same query under two configurations that differ only in the utility definition: the default binary-utility Pareto-500 and the ablation with $\tau = 0.05$ soft-sigmoid

smoothing (Section 4.6). Both decompose into the same two facets and both report `coverage = 1.0`. The selections diverge on three pids: the binary run packs *Port Blair* (`context_13`), *Pangi Territory* (`context_18`), and *WaterColor, Walton County, Florida* (`context_19`)—passages with generic "county/headquarters" lexical overlap whose underlying verifier scores sit close enough to $\bar{\alpha}_f$ that the boolean utility ties them with the true target. The greedy solver breaks the tie by cost/order and lands on Walton County; Llama-3-8B-Instruct, given no Maryland-relevant passage, returns an empty answer. Under $\tau = 0.05$, the same near-threshold verifier scores are preserved as smoothed marginal utilities that separate the candidates: `context_7` (*Hughesville is a census-designated place in Charles County, Maryland*) edges out the Florida/India/Congo distractors because its score aligns more strongly with the "Southern Maryland" bridge. The packed set swaps `context_13`, `context_18`, `context_19` for `context_0`, `context_9`, `context_7`; the LLM returns *"Charles County"* (EM match). Evidence tokens drop from 359 to 294 and latency from 521 to 344 ms. This is the concrete mechanism behind the ablation: smoothing does not change the semantic definition of coverage, but it prevents the greedy packer from wasting its budget on ties-by-accident when multiple candidates sit at $p \approx \bar{\alpha}_f$.

Vignette C — Top-k retrieval vs. TRIDENT facet coverage (same query, side-by-side). We extend the running example of Figure 7—*"Who is the mother of the director of Polish-Russian War?"*—into a side-by-side comparison of two selection rules over the same retrieved candidate pool. Both rules see the same 20-passage pool returned by the shared retriever; they differ only in how they choose which passages to keep.

Top-k retrieval ($k = 3$, ranked by similarity to the full query string). The pool's three top-ranked passages are: p_1 (the *Polish-Russian War* film synopsis, mentioning the title heavily), p_2 (a 1792 Polish–Russian war article matching "Polish-Russian War" lexically), p_5 (a war-history overview that mentions the film in a single sentence). None of the three names the director, and none names the director's mother. The LLM, given $\{p_1, p_2, p_5\}$, returns either an unsupported guess or a refusal; provenance check fails because no selected passage names *Małgorzata Braunek*. EvTok consumed: $\sim$580. Outcome: incorrect answer with no certificate.

TRIDENT facet coverage (Safe-Cover, $\alpha_{query} = 0.10$, $|F_{struct}| = 2$). The miner emits three facets: ENTITY: *Polish-Russian War* (optional, $\bar{\alpha}_f = 1$); RELATION DIRECTOR: *director-of(Polish-Russian War, ?x)* (structural, $\bar{\alpha}_f = 0.05/T_f$); RELATION MOTHER: *mother-of(?x, ?y)* (structural). Per-facet shortlisting promotes p_4 (the film infobox, naming *Xawery Żuławski*) to the top of the DIRECTOR list (verifier score above $\bar{\alpha}_f$), and the Hop-2 pass conditioned on that binding promotes p_7 (the director's biography, naming *Małgorzata Braunek*) to the top of the MOTHER list. The greedy min-cost solver selects $\{p_4, p_7\}$. Both structural facets are certified; the optional ENTITY facet contributes utility but no FWER weight. EvTok consumed: $\sim$310. Outcome: *Małgorzata Braunek* (EM match) with one structural certificate per hop.

Table 29 summarises the divergence. Top-k retrieval optimises whole-query similarity and packs three passages with strong lexical anchor on *Polish-Russian War* but no path to the answer; TRIDENT shortlists each facet against its own typed key, so the selected set is the union of facet-specific top-T_f's rather than a single global top-k. The mechanism is not stronger retrieval—both rules see the same pool—but stronger selection: "coverage" is a property of which structural requirements the selected set jointly supports, not of whole-query similarity.

R Proofs

Proof of Proposition 1 (anti-conservativeness of pre-selection calibration). Fix bin b and a null pair (p, f). Let F_0 be the unconditional null cdf of $s(p, f)$ in bin b and F_0^{sel} its cdf conditional on $\{(p, f) \in \text{Shortlist}(f)\}$. By assumption $F_0^{\text{sel}} \leq F_0$ pointwise, i.e., the conditional null score distribution first-order stochastically dominates the unconditional null. (A score-monotone top-T_f rule is the canonical sufficient condition; mere non-negative correlation between the shortlister's ranking signal and $s(p, f)$ is *not* sufficient in general for stochastic dominance, so we take SD as the primitive assumption rather than as a corollary.) The pre-selection p-value is $\tilde{\pi}(p, f) = 1 - F_0(s(p, f))$. For $\tau \in [0, 1]$,

$$\Pr_{F_0^{\text{sel}}}\big[\,\tilde{\pi}(p,f) \leq \tau\,\big] = \Pr_{F_0^{\text{sel}}}\big[\,F_0(s(p,f)) \geq 1-\tau\,\big] = 1-F_0^{\text{sel}}\big(F_0^{-1}(1-\tau)\big) \geq 1-F_0\big(F_0^{-1}(1-\tau)\big) = \tau,$$

Table 29: Top-k retrieval vs. TRIDENT facet coverage on the same retrieved pool for *"Who is the mother of the director of Polish-Russian War?"*. Both rules consume the same 20-passage candidate pool returned by the shared retriever and the same verifier; they differ only in selection logic. "Cover DIRECTOR" / "Cover MOTHER" indicates whether the selected set contains a passage that the verifier scores above $\bar{\alpha}_f$ for that structural facet.

Aspect	**Top-k retrieval (k=3)**	**TRIDENT Safe-Cover**
Selection unit	Whole query	Per typed facet
Selected passages	$\{p_1, p_2, p_5\}$	$\{p_4, p_7\}$
Cover DIRECTOR	No	Yes (cert. on p_4)
Cover MOTHER	No	Yes (cert. on p_7, Hop-2)
Provenance for answer	Fails	Passes (cited p_7)
EvTok	$\sim$580	$\sim$310
Answer	Unsupported / refusal	*Małgorzata Braunek* (EM)
Certificate	—	2 structural (FWER ≤ 0.10)

where the inequality uses $F_0^{\text{sel}} \leq F_0$ pointwise. Strict inequality holds whenever $F_0^{\text{sel}} \neq F_0$ at the threshold $F_0^{-1}(1-\tau)$, i.e., whenever the conditioning is informative for that threshold. For discrete verifier scores, replace F_0^{-1} by the generalized quantile $F_0^{-1}(u) = \inf\{x : F_0(x) \geq u\}$ and read F_0 as the left-continuous cdf at the threshold; the same chain of inequalities holds, and strict inflation occurs at thresholds where the selection-conditional null places additional mass above the unconditional quantile. Hence the pre-selection p-value rejects more often than the nominal τ on the shortlisted subset. □

Proof of Theorem 1 (marginal validity). Fix a bin b and a null pair (p, f), so $s(p, f)$ is exchangeable with the calibration scores $\{s(u) : u \in \mathcal{N}_b\}$ conditional on b. Let $N = |\mathcal{N}_b|$ and consider the augmented score vector of size $N + 1$. By exchangeability, the rank of $s(p, f)$ among the $N + 1$ scores (with ties broken by independent uniform randomization) is uniformly distributed on $\{1, 2, \ldots, N+1\}$. The conformal p-value in (1) equals $R/(N+1)$ where R is that rank from the top. Hence $\Pr[\pi(p, f) \leq \tau \mid b] = \Pr[R \leq \tau(N+1) \mid b] = \lfloor \tau(N+1) \rfloor / (N+1) \leq \tau$. Marginalizing over b yields the claim. □

Proof of Theorem 2 (query-level FWER). Let E_f be the event that any null shortlisted pair for structural facet $f \in F_{\text{struct}}(q)$ is certified, i.e., $\pi(p, f) \leq \bar{\alpha}_f$ under the null for some $p \in \text{Shortlist}(f)$. A false structural facet-support certificate is emitted within the query iff $\bigcup_{f \in F_{\text{struct}}(q)} E_f$ occurs. By Theorem 1, each per-pair certification event has probability at most $\bar{\alpha}_f = \alpha_{\text{query}} / (|F_{\text{struct}}(q)| \cdot T_f)$. Applying the union bound and substituting $\bar{\alpha}_f$,

$$\begin{aligned} \Pr\left[\bigcup_{f \in F_{\text{struct}}(q)} E_f \right] &\leq \sum_{f \in F_{\text{struct}}(q)} \sum_{j=1}^{T_f} \Pr[\pi_{fj} \leq \bar{\alpha}_f] \\ &\leq \sum_{f \in F_{\text{struct}}(q)} T_f \cdot \frac{\alpha_{\text{query}}}{|F_{\text{struct}}(q)| \cdot T_f} \;=\; \alpha_{\text{query}}. \end{aligned}$$

Optional facets ($\bar{\alpha}_f = 1$) are excluded from the event by construction and do not enter the bound. Non-null pairs contribute no false certificates, so the bound extends unchanged. □

Proof of Theorem 3 (replay validity). Safe-Cover is a composition of pure functions of the episode contract E: (i) the retriever, shortlister, and verifier are deterministic given their weights, inputs, and logged tie-break seed; (ii) the p-value map (1) is a rational function of $s(p, f)$ and $\{s(u) : u \in \mathcal{N}_b\}$, with score ties broken by a logged RNG seed; (iii) the thresholds $\{\bar{\alpha}_f\}$ are arithmetic functions of $(\alpha_{\text{query}}, |F_{\text{struct}}(q)|, T_f)$; (iv) the greedy min-cost solver in (3) iterates over passages in a total order fixed by shortlister tie-breaks and selects by a deterministic marginal-coverage-per-cost rule; (v) the canonical certificate payload serializes fields in a fixed schema with sorted keys and stable numeric formatting. Bit-identical inputs under a bit-identical composition of pure functions yield a bit-identical output, so $\mathcal{C}'(q, E) = \mathcal{C}(q, E)$ and $\text{SHA-256}(\mathcal{C}') = \text{SHA-256}(\mathcal{C})$. □

Proof-sketch summary. Both the test-time path and the calibration path execute the *same* shortlister logged in the contract E, so the test pair's score and the held-out null scores share a common selection-conditional null within each Mondrian bin b. Conditional exchangeability gives marginal validity of the conformal p-value (Theorem 1); a Bonferroni union bound over the at most $|F_{\text{struct}}(q)| \cdot T_f$ shortlisted hypotheses per query yields the query-level FWER (Theorem 2). Theorem 3 then states that any other replay under the same E produces a bit-identical certificate.

S Contract analyses

S.1 Deployment coverage ledger

We reconstruct complete Bonferroni cover from split-matched deployment logs. All $5{,}516$ issued LLM-miner records are structural declarations.

Table 30: LLM-miner coverage at $\alpha_q = 0.10$: count / denominator (%). Complete cover certifies all mined structural facets.

Dataset	Contract	≥ 1 structural declaration	Certified passage retained	Complete certified output
HotpotQA	Lex	540/7,405 (7.29)	539/7,405 (7.28)	20/7,405 (0.27)
HotpotQA	NLI	653/7,405 (8.82)	651/7,405 (8.79)	24/7,405 (0.32)
2Wiki	Lex	1,200/12,576 (9.54)	1,196/12,576 (9.51)	63/12,576 (0.50)
2Wiki	NLI	2,115/12,576 (16.82)	2,112/12,576 (16.79)	74/12,576 (0.59)
MuSiQue	Lex	243/2,417 (10.05)	241/2,417 (9.97)	0/2,417 (0)
MuSiQue	NLI	66/2,417 (2.73)	65/2,417 (2.69)	0/2,417 (0)

S.2 Calibration-only controls and adaptive replay

Calibration-only controls fix queries, candidates, selected hypotheses, verifier, and nominal level: three datasets, six selectors, two verifiers, three levels, and five splits give 108 configurations (540 evaluations). Selectors are lexical, dense, cross-encoder, verifier-based, fitted-threshold, and dense-to-cross-encoder cascade. Table 31 counts point estimates at or below nominal; conformal validity assumes exchangeability.

Table 31: Calibration controls: aggregate and all-five-split counts.

Calibration	Configurations	All five splits
Replay conformal	108/108	108/108
Pre-selection conformal	82/108	44/108
Replayed empirical quantiles	108/108	103/108
Isotonic	107/108	54/108
Raw-score / Platt thresholds	58/108	16/108

Including a fixed shuffled control, replay at $\alpha = 0.05$ meets nominal in $42/42$ cells ($210/210$ query-grouped splits), versus $26/42$ pre-selection aggregates.

Adaptive title binding changes 80.6–87.4% of ordered stage-two top-five families. Full-path replay repeats selection, binding extraction, hypothesis construction, and shortlisting; false-event rates at $\alpha = 0.05$ are 1.95–2.64% across six settings; pre-selection exceeds 5% in two, reaching 7.07% and 8.58% on MuSiQue.

At pre-specified 20% issuance on identical held-out queries/candidates, replay retains annotated support on $150/290$ queries (51.72%), versus raw NLI's $136/290$ (46.90%): paired $+4.83$ points, 95% CI $[1.03, 9.31]$.

S.3 Miner coverage and verifier sensitivity

Two independently prompted temperature-zero judges assess facet–hop alignment on 900 queries each from 2Wiki/MuSiQue. Both-judge represented-hop rates are 78.67% $[76.45, 80.81]$ and 83.80%

$[81.74, 85.77]$; every annotated hop is represented on 66.44% and 73.44% of queries. Missing-hop rates are $21.33\%/16.20\%$ ($33.56\%/26.56\%$ of queries miss a hop). Shuffled-facet representation is $0.76\%/0.17\%$.

After rebuilding, reference/alternative-verifier false-event rates at $\alpha = 0.10$ are $5.67/4.67\%$ (HotpotQA), $2.00/2.40\%$ (2Wiki), and $8.33/7.17\%$ (MuSiQue). Declaration-set overlap is 11.70–18.52%.

S.4 Contract maintenance and validity–power trade-offs

Re-mining costs 3.35–4.40 aggregate worker-seconds per query. Once facets are available, loading/scoring $7{,}166$–$31{,}201$ pairs takes 0.2–1.5 minutes.

Query-level drift at $\alpha = 0.10$ gives matched rates 5.08%, 7.84%, and 6.35% on 2Wiki, HotpotQA, and MuSiQue. Stale 2Wiki/MuSiQue calibration on HotpotQA gives $19.46\%/15.43\%$.

Across selector policies at $\alpha = 0.10$, automatic at-least-one benchmark-support yield is 9.56–12.00% (HotpotQA), 10.72–17.22% (2Wiki), and 4.61–8.00% (MuSiQue), with false-event means 2.89–6.28%. At pre-specified $\alpha = 0.20$, yield increases to 16.50–21.17%, 21.67–25.33%, and 8.61–14.89%, with false-event means 5.72–9.72%.

S.5 Learned compression and native-system baseline settings

RECOMP [Xu et al., 2024] shares the full splits, candidates, answerer, prompt, decoding, and 500-token cap; we use the released checkpoint without evaluation-set tuning.

Table 32: Full-split RECOMP at 500 tokens: F1 (%) / evidence tokens.

Method	HotpotQA	2Wiki	MuSiQue
RECOMP	59.87 / 492.11	41.37 / 490.34	21.70 / 480.10

Search-R1 uses released Qwen2.5-7B PPO and up to five top-three search turns (Table 6). HippoRAG2 [Gutiérrez et al., 2025] uses native OpenIE graph/dense retrieval with Llama-3-8B-Instruct.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) state four contributions: (i) post-selection evidence accountability for RAG, formalizing support verification as selected-hypothesis testing under a frozen episode contract (Proposition 1; §4.2); (ii) a frozen passage–facet table $T(q)$ as a shared replayable representation across regimes; (iii) Safe-Cover, which composes replay-calibrated conformal p-values with a budgeted min-cost cover to control query-level Bonferroni FWER over false structural support certificates, with audited false-certificate point estimates below the nominal α_{query}=0.10 under rebuilt LLM-miner conformal pools (Theorems 1, 2); and (iv) Pareto-Knapsack, the efficiency regime, which packs evidence by marginal facet gain per token under B_{ctx} and obtains mean multi-hop F1 41.74 versus RECOMP's 40.98 at 408 versus 487.52 mean evidence tokens under the matched 500-token protocol. The Safe-Cover claims are theorem-backed and contract-scoped; the Pareto claims are empirical quality–cost claims with no FWER guarantee. The exact theorem-by-theorem boundary is in Appendix D (Tables 11, 12).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A concise limitations discussion appears in the main paper in Section 4.6 ("Certificate power"; contract-dependence and mechanism ablations), covering miner/verifier-bound certificate yield and precision, the contract-scoped nature of the FWER bound, and that selection-conditional calibration is Safe-Cover's validity-recovery ingredient rather than Pareto's F1 driver. The full statement is in Appendix A, which adds drift-stress quantification (Appendix H), single-passage conservatism on diffuse MuSiQue evidence, the regime where compression baselines outperform TRIDENT (KILT-FEVER), the observational answer-linked slice analysis, native-system baseline settings, and the scope of the selection-baseline comparison.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3 states all formal results with explicit assumptions: Proposition 1 (anti-conservativeness of pre-selection calibration under stochastic dominance of the selection-conditional null), Theorem 1 (marginal validity of the conformal p-value under conditional exchangeability within a Mondrian bin), Theorem 2 (query-level Bonferroni FWER over false structural-facet certificates), and Theorem 3 (bit-identical replay under contract). Full proofs are in Appendix R. A theorem-by-theorem statement of what is and is not covered, including which empirical claims carry no theorem, appears in Appendix D (Tables 11, 12). The validity claim is explicitly scoped to the logged episode contract throughout.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 specifies datasets (HotpotQA, 2WikiMultiHopQA, MuSiQue, KILT-FEVER as a single-claim boundary case), LLMs (Llama-3-8B-Instruct, Qwen3-8B), decoding (T=0), evidence-token caps, and the two deployed LLM-miner contracts (LLM+lex, LLM+NLI) with contract-rebuilt conformal pools. Appendix C (Table 9) lists every episode-contract field required for bit-identical replay (Theorem 3). Appendix E covers truncation policy, token accounting (EvTok vs. TotTok), latency definition, and compute environment. Appendix M documents baseline parity and budget enforcement against LLMLingua, Sentence-Compress, BGE-Rerank, LLM-Select(+NLI), and NLI-Filter. Mondrian bin sensitivity (L, $N_{\min}$) is in Appendix O; cross-corpus drift in Appendix H; contract-change controls and the RECOMP protocol in Appendix S. Code is released at `https://github.com/tomzhang97/Trident`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is released at `https://github.com/tomzhang97/Trident`. The three primary multi-hop evaluation datasets (HotpotQA, 2WikiMultiHopQA, MuSiQue) and the KILT-FEVER boundary case are publicly available. The paper provides full algorithmic specifications: Algorithms 1 (Safe-Cover) and 2 (Pareto-Knapsack), the conformal p-value (Eq. 1), the Bonferroni allocation (Eq. 2), the cover objective (Eq. 3), the heuristic-mode floor (Eq. 5), the contract-field schema (Table 9), and bin-merging / abstention reason codes in Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies LLMs (Llama-3-8B-Instruct, Qwen3-8B), decoding (temperature 0), evidence-token caps (Pareto-400/500/1500; Safe-Cover at B_{ctx}=2000 and 4000), the query-level error budget (α_{query}=0.10), and the two deployed LLM-miner contracts (LLM+lex, LLM+NLI). Appendix E adds truncation policy, token accounting, latency definition, and compute environment; Appendix M documents per-baseline parity and budget enforcement. Conformal pools are rebuilt separately for each deployed contract on a 50/50 calibration/audit split (Appendix K).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports 95% intervals for audit-pool false-certificate rates. Table 23 reports bootstrap 95% CIs (B=10,000) for answer-linked versus no-cert EM/F1 differences. Table 13 reports Wilson intervals for EM and stratified-bootstrap intervals (B=1,000) for F1 in the system comparisons with interval logs; the Pareto-500 versus VanillaRAG comparison is discussed in Appendix E. Table 21 reports Wilson 95% CIs for offline LLM-miner rates. The RECOMP comparison reports point estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on a local workstation with $4\times$ NVIDIA RTX 4090 (24 GB) and $3\times$ NVIDIA RTX Pro 6000 (96 GB) GPUs. Individual end-to-end runs ranged from a few hours to roughly 20 hours of wall-clock time depending on dataset, LLM (Llama-3-8B-Instruct vs. Qwen3-8B), and selection regime (Pareto-Knapsack vs. Safe-Cover); the full project required additional compute for miner/shortlister sweeps and conformal-pool calibration that did not appear in the final paper. Per-query end-to-end latency under fixed batching and caching policies is reported alongside every method and configuration in Tables 5 and 6, with the latency definition in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and the work conforms in every respect. The research uses publicly available QA benchmarks, open-weight language models, and proposes a methodological framework for evidence selection. No ethical concerns are identified.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: TRIDENT is a methodological framework for post-selection evidence accountability in RAG. By calibrating support tests under the policy that selected the evidence, it tightens the auditability of citation claims and provides a typed-abstention path when certification is infeasible. Direct negative societal impacts beyond those inherent to the underlying language models are not foreseen and are outside the scope of this work.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, datasets, or other assets with high misuse risk. TRIDENT is a selection and certification framework that operates on top of existing retrieval and generation components.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models are cited with their original publications and used under their stated licenses: HotpotQA (CC BY-SA 4.0), 2WikiMultiHopQA (Apache 2.0), MuSiQue (CC BY 4.0), KILT (MIT) wrapping FEVER (CC BY-SA 3.0), Llama-3-8B-Instruct (Meta Llama 3 Community License), and Qwen3-8B (Apache 2.0). Baseline methods, models, and checkpoints (RECOMP, LLMLingua, BGE reranker, DeBERTa-v3, HippoRAG2, IRCoT, and Search-R1 with its released Qwen2.5-7B-Instruct checkpoint) are cited to their source publications and used under their released licenses. Use is restricted to non-commercial academic research, consistent with each license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the TRIDENT code, released at `https://github.com/tomzhang97/Trident`. The paper does not release new datasets or pre-trained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects. All evaluations are automated using existing QA benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs (Llama-3-8B-Instruct, Qwen3-8B) are core components of the RAG pipeline evaluated in this work. Llama-3-8B-Instruct is used in three logged roles within the episode contract: (i) as the contract-logged facet planner / miner (Section 3.1; LLM-miner ablation in Appendix K); (ii) as the LLM-Select baseline that scores candidate passages; and (iii) as the answering LLM under temperature-0 decoding. Qwen3-8B is reported as a swap-in LLM with retrieval/verifier/calibration/cap held fixed (Section 4.1, Tables 5, 6). The reference verifier is a DeBERTa-v3 entailment scorer. The facet–hop alignment diagnostic uses two independently prompted temperature-zero LLM judges (Appendix S.3). Prompt, decoding seed, and JSON-normalization rules for the LLM facet planner are frozen in the episode contract (Table 9).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.