PACKETCERT: SET-LEVEL EVIDENCE CERTIFICATION FOR COMPACT RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Compact RAG makes an irreversible decision before generation: which retrieved passages survive the context budget. Passage-wise selection can preserve answer F1 while dropping a required hop, the qualifier governing a claim, or counterevidence already present in the pool. PACKETCERT certifies the selected set by compiling a query–pool pair into linked proposition, qualifier, and available-counterevidence requirements; learning coverage and selection scores from exact typed-cover targets; decoding witnessed sets under a hard cap; and using a separately parameterized verifier for deficiency-directed expansion or abstention. Independent packet judgments provide the primary validation. On a 600-packet cohort balanced over HotpotQA, MuSiQue, 2WikiMultiHopQA, and ASQA, approval precision/recall is 91.4/88.7%, versus 82.5/77.9% for an identical-safeguard CE+NLI selector; the certificate score reaches AUROC 0.881 and 2.9% harmful-omission risk at 80% coverage, compared with 7.6% for SURE-RAG. On HotpotQA, compiled-requirement coverage is 92.8% at 1.4K evidence tokens, versus 81.7% at 2.1K for CE+NLI. In a prespecified analysis of 36 answer-F1 near ties, approval reverses 14 choices and blinded judges prefer its winner in 11. On all 948 ASQA examples, PACKETCERT attains 33.3 Disambig-F1 with 5.9% harmful omission at 1.6K tokens. Set-level certification therefore improves compact evidence construction and supplies a human-aligned tie-breaker where answer quality cannot distinguish operating points.

1 INTRODUCTION

The context handed to a RAG reader is a set, but most compact selectors score its members independently. This mismatch is consequential: two packets can support the same answer string while only one contains every reasoning hop, keeps a temporal or entity qualifier attached to its claim, and exposes a retrieved caveat that changes the supported answer. After selection, neither the reader nor a citation metric can recover omitted text. Answer F1 can therefore tie operating points that present materially different evidence.

A *typed evidence certificate* makes that set structure explicit. Proposition nodes state what the packet must jointly establish, qualifier nodes retain their parent links, and counterevidence nodes activate for answer-changing evidence found in the fixed pool. Each node carries a passage witness and threshold margin; a separately parameterized set verifier checks joint sufficiency. The resulting controller has three meaningful terminal actions under budget: certify, expand toward named deficient nodes, or return UNCERTIFIEDUNDERBUDGET.

PACKETCERT constructs this object rather than scoring a completed answer. A query–pool compiler builds the linked graph; a frozen DeBERTa coverage head supplies requirement–passage edges; a separate selection head learns from exact typed-cover membership; and a budgeted decoder combines learned selection probability with marginal witnessed coverage per token. The verifier receives the unordered packet and linked requirements, never selector logits. Every verification call is preceded by the hard single-witness test, and rejected packets are expanded using the verifier's deficient requirement IDs.

The evidence supports three levels of the claim. First, independent labels on 600 packets show that certificate scores predict adequacy (AUROC 0.881) and reduce Risk@80 to 2.9%. Second, against CE+NLI with the same counterevidence safeguard, PACKETCERT raises compiled-requirement coverage by 11.1 points and approval precision by 8.9 points while using fewer evidence tokens.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Third, among 36 prespecified answer-F1 near ties, certificate approval agrees more strongly with blinded preference ($\tau = 0.64$ versus 0.29); ASQA confirms that lower omission need not trade away answer quality.

The paper contributes:

- **Certificate synthesis:** a relation-aware set objective connects propositions, attached qualifiers, and available counterevidence to explicit witnesses, margins, token feasibility, and a set verdict.
- **Budgeted construction:** exact typed-cover supervision, a learned selection prior, hard witnessed decoding, verifier-directed expansion, and abstention operate on the same stable requirement IDs.
- **Independent decision evidence:** blinded adequacy and preference judgments, identical-safeguard controls, ablations, corrected requirements, verifier changes, clutter, budget sweeps, and ASQA test both mechanism and consequence.

Together, these contributions turn evidence adequacy into both a compact-construction target and a human-aligned criterion for resolving answer-F1 near ties.

2 Related Work

2.1 Retrieval, compact selection, and compression

Dense retrieval, late interaction, cross-encoder reranking, and iterative retrieval improve candidate recall or order (Karpukhin et al., 2020; Khattab & Zaharia, 2020; Santhanam et al., 2022; Nogueira & Cho, 2019; Izacard et al., 2022; Trivedi et al., 2023). PacketCert addresses the next, irreversible handoff: every fixed-pool constructor sees the same top-K, so available-but-omitted evidence is attributable to set construction. Iterative Retrieval is reported as adaptive-retrieval context, not as a matched fixed-pool control.

Prompt compression reduces reader cost after evidence is chosen (Jiang et al., 2023; Xu et al., 2024), but can preserve answer quality while degrading citation grounding (Li et al., 2026). ECoRAG adaptively retains answer-supporting evidence (Jeong et al., 2025); CompSelect learns compact clues (Zhang et al., 2026). Both are strong shared-pool baselines. PacketCert differs in the construction invariant: the final set must jointly witness linked proposition, qualifier, and available-counterevidence requirements before generation.

2.2 RAG evaluation and sufficiency

RAGAS, ARES, and RAGChecker score context relevance, answer faithfulness, or claim-level pipeline quality (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024); they are comparators in the selective-risk study. Sufficient Context supports selective generation (Joren et al., 2025), while SURE-RAG predicts support, refutation, or insufficiency for a *candidate answer* (Qiu et al., 2026). Entropic Claim Resolution acquires claims to reduce answer uncertainty (Di Gioia, 2026). In contrast, PacketCert compiles answer-free obligations from the query and pool and constructs the bounded packet that satisfies them.

2.3 Set cover, attribution, and contradiction handling

Citation and factuality systems judge generated claims (Menick et al., 2022; Fabbri et al., 2022; Min et al., 2023; Gao et al., 2023); a certificate instead constrains the evidence available before generation. Classical set cover supplies the combinatorial backbone (Chvátal, 1979; Feige, 1998). The new object is a linked semantic cover with hard witnesses, a learned passage prior, global verification, diagnosed expansion, and abstention. The identical-safeguard CE+NLI control isolates this set coupling from contradiction detection alone.

3 TYPED EVIDENCE CERTIFICATE SYNTHESIS

3.1 OBJECTS, REQUIREMENTS, AND THREAT MODEL

For query q, retrieval returns the immutable pool $\mathcal{P}_q = \{p_1, \ldots, p_K\}$. Construction chooses $\mathcal{S} \subseteq \mathcal{P}_q$ under token budget B. Requirements are typed:

$$\mathcal{R}_q = \mathcal{R}_q^{\text{prop}} \cup \mathcal{R}_q^{\text{qual}} \cup \mathcal{R}_q^{\text{counter}}.$$

Proposition and qualifier candidates are induced from the query, then grounded against the fixed pool. Counterevidence candidates are induced without an answer: an NLI model scans pool passages against the grounded proposition/qualifier statements and creates an active counter slot when a passage exceeds the development-tuned activation threshold η_{act}. The separate coverage threshold τ_c governs whether the selected packet witnesses that active slot. Thus the certificate can require exposure of answer-changing evidence without presupposing a generated answer.

Requirements form a directed graph rather than an unstructured list. Every qualifier stores the identifier of the proposition it scopes, and every counterevidence slot stores the proposition or qualifier relation it contests. The text scored in Eq. 1 serializes both the dependent requirement and its parent statement; a passage cannot satisfy a detached qualifier phrase. Different requirements may use different witnesses, but the verifier receives the complete relation graph and unordered witness set, and rejects combinations whose individually covered statements are jointly incompatible. Appendix A gives the serialization and a worked certificate.

The stressor may exploit semantic redundancy, partial multi-hop support, distractor similarity, or budget pressure to make an incomplete packet appear relevant. The protocol holds retrieval pools and benchmark support annotations constant across constructors; a separately parameterized verifier evaluates each proposed packet after selection. For passage p_i and requirement r_j, the pinned coverage model is

$$v_{ij} = \sigma(\text{CrossEnc}(p_i, r_j)). \tag{1}$$

Thresholds $\tau_{\text{prop}}, \tau_{\text{qual}}, \tau_c$ are selected on held-out training calibration partitions and frozen before evaluating any reported split.

3.2 CERTIFICATE OBJECT AND APPROVAL PREDICATE

Let $c_j(\mathcal{S}) = \max_{p_i \in \mathcal{S}} v_{ij}$, with $c_j(\emptyset) = 0$, and let $T(\mathcal{S})$ be token cost. Each requirement receives witness $\pi(j) = \arg\max_{p_i \in \mathcal{S}} v_{ij}$ and margin $\delta_j = c_j(\mathcal{S}) - \tau_{t(j)}$. The separately parameterized verifier returns sufficiency score $\nu(\mathcal{S}, q) \in [0, 1]$, verdict $V = \mathbb{1}[\nu \geq \tau_V]$, and, on rejection, a set $D_V \subseteq \mathcal{R}_q$ of deficient requirements. A typed certificate is

$$\mathcal{C}_B(q) = \big(\mathcal{S}, \pi, \boldsymbol{\delta}, \nu\big).$$

Its complete approval predicate is

$$\begin{aligned} A_B(\mathcal{S}, q) = \mathbb{1}\big[T(\mathcal{S}) \leq B \ \wedge \ \min_{r_j \in \mathcal{R}_q^{\text{prop}}} c_j(\mathcal{S}) \geq \tau_{\text{prop}} \\ \wedge \ \min_{r_j \in \mathcal{R}_q^{\text{qual}}} c_j(\mathcal{S}) \geq \tau_{\text{qual}} \ \wedge \ \min_{r_j \in \mathcal{R}_q^{\text{counter}}} c_j(\mathcal{S}) \geq \tau_c \ \wedge \ \nu(\mathcal{S}, q) \geq \tau_V\big]. \end{aligned} \tag{2}$$

Empty qualifier or counterevidence types make their conjunctions true; proposition requirements are nonempty. For selective evaluation, we form the minimum normalized certificate margin

$$\rho(\mathcal{S}, q) = \min\left\{ \frac{\nu(\mathcal{S}, q)}{\tau_V}, \min_{r_j \in \mathcal{R}_q} \frac{c_j(\mathcal{S})}{\tau_{t(j)}} \right\}$$

and map it to approval score $s_B = \text{IsoCal}(\rho)$ using isotonic calibration fitted on the held-out training calibration partition. The map and all thresholds are frozen before evaluation; adequate packets are the positive class for AUROC, AUPRC, ECE, and risk–coverage analysis.

The ideal certificate-synthesis target is

$$\mathcal{S}^* = \arg\min_{\mathcal{S} \subseteq \mathcal{P}_q} T(\mathcal{S}) \quad \text{s.t.} \quad A_B(\mathcal{S}, q) = 1. \tag{3}$$

The online controller approximates this target by typed-cover construction followed by verifier-guided expansion. A return of CERTIFIED is one-sided and checkable: every hard witness, the budget test, and the verifier threshold in Eq. 2 must pass. Budget exhaustion returns UNCERTIFIEDUNDERBUDGET. Operational sufficiency can admit several certificates, so Eq. 3 targets the lowest-cost approved set rather than a unique explanation.

3.3 TRAINING AND COMBINATORIAL STRUCTURE

Training is staged to keep coverage validity distinct from passage selection. The requirement–passage encoder and coverage head in Eq. 1 are first fitted on training-only witness pairs, calibrated, and then frozen. The offline teacher binarizes these frozen edges as $a_{ij} = \mathbb{1}[v_{ij} \geq \tau_{t(j)}]$ and solves exact weighted set cover. A separate selector head reads the query, all linked requirement records, passage p_i, and normalized token cost to produce s_i, with $z_i = \sigma(s_i/T_s)$. During selector training, gradients flow through z_i and the selector head but not through the frozen v_{ij}. Differentiable noisy-OR coverage is

$$\widetilde{c}_j = 1 - \prod_i (1 - z_i v_{ij}),$$

which is zero when no passage is selected and lies in $[0, 1]$. The objective is

$$\mathcal{L}_{\text{select}} = \sum_j w_j \,\text{softplus}\big(\tau_{t(j)} - \widetilde{c}_j\big) + \lambda_1 \sum_i z_i |p_i|_{\text{tok}} + \lambda_2 \sum_{r_j \in \mathcal{R}_q^{\text{counter}}} \text{softplus}(\tau_c - \widetilde{c}_j) + \lambda_3 \mathcal{L}_{\text{ILP}}. \tag{4}$$

$\mathcal{L}_{\text{ILP}}$ is binary cross-entropy from z_i to exact teacher membership, type weights w_j compensate training-frequency imbalance, and T_s is annealed. At inference, $\widehat{z}_i = \sigma(s_i)$ is the learned passage prior used in Eq. 5; requirement gains use unit weights and passage costs remain strictly positive.

Noisy-OR shapes selection but cannot certify a packet: several weak edges can increase $\widetilde{c}_j$, whereas Eq. 2 requires a single passing witness through c_j. Every decoded set is therefore rescored against the frozen hard edges before verification.

The exact teacher and the online controller solve different problems. Classical unbudgeted, unmodulated weighted greedy cover has the H_m approximation bound (Chvátal, 1979); the deployed decoder additionally applies a learned prior, remaining-budget feasibility, and verifier-directed expansion. Its formal invariant is certification soundness under Eq. 2, detailed in Appendix C.

4 PACKETCERT: CERTIFICATE CONSTRUCTION

PACKETCERT synthesizes a typed certificate in four coupled stages. First, Llama-3.1-70B-Instruct emits proposition and qualifier slots with stable IDs and parent links; pool-conditioned NLI discovery adds caveat and contradiction slots linked to the statement they contest. Second, the frozen DeBERTa coverage head evaluates Eq. 1, while the separately trained selector head supplies $\widehat{z}_i$. Third, the constructor repeatedly selects

$$p^* = \arg\max_{p_i \in F(\mathcal{S})} \frac{\widehat{z}_i \Big(\sum_{j \in U(\mathcal{S})} a_{ij} + \sum_{j \in D_V} v_{ij} \Big)}{|p_i|_{\text{tok}}}, \tag{5}$$

where $F(\mathcal{S}) = \{p_i \notin \mathcal{S} : T(\mathcal{S}) + |p_i|_{\text{tok}} \leq B\}$, $U(\mathcal{S})$ contains uncovered requirements, and D_V is empty before verification. Thus the ILP-trained prior changes the online order while hard edges determine whether coverage is complete. Fourth, the construction verifier receives the query, linked records, and unordered packet and returns a score, verdict, and deficient IDs. Rejection populates D_V, directing Eq. 5 toward corroboration for diagnosed obligations; absent a valid ID, the lowest normalized-margin requirement is used. Schema-invalid output rejects. Approval or exhaustion of feasible positive-gain candidates terminates the controller. Gurobi is used only to generate training targets. On development data, the online constructor retains 96.2% of teacher coverage at $18\times$ faster inference. Algorithm 1, the operational schemas in Appendix A, and the certificate-evidence map in Appendix B expose the controller from induced requirement to terminal decision.

5 EXPERIMENTS

5.1 RESEARCH QUESTIONS AND PROTOCOL

The experiments separate validity, mechanism, stability, and transfer. RQ1 tests certificate scores against blinded packet adequacy and preference. RQ2 isolates construction from the shared NLI safeguard and strong compact selectors. RQ3 perturbs requirement labels, verifier family, prompt order, clutter, and budget. RQ4 tests long-form answer quality, omission, latency, and cost on ASQA.

Datasets. Evaluation uses the public development splits of HotpotQA (7,405 questions), MuSiQue (2,417), 2WikiMultiHopQA (12,576), and ASQA (948) (Yang et al., 2018; Trivedi et al., 2022; Ho et al., 2020; Stelmakh et al., 2022). ASQA exposes multiple facets and dense citations without injected clutter. HotpotQA-Hard expands each top-K=50 pool to K=80 with matched distractors and NLI-flagged partial counterevidence (37% and 5% of passages on average).

Systems and controls. Top-k, MMR, CE Rerank, CE+NLI, RECOMP, LLMLingua, ECoRAG, CompSelect, and PACKETCERT receive the identical Contriever-v2 top-50 pool and $B = 2.1$K-token maximum in Table 1. CE+NLI inserts NLI-active passages in descending activation score, then fills by cross-encoder rank; any whole passage exceeding the residual budget is skipped. It shares PACKETCERT's detector, activation threshold, and keep rule, but has no requirement graph or set verification. ECoRAG and CompSelect use their published scores and stopping rules over the common pool; all methods retain whole passages and skip overflow. Their native gates never call PACKETCERT's verifier. Iterative Retrieval and Long-Context are adaptive/high-token references. The GPT-4o pass rate is a controller diagnostic because PACKETCERT uses that verifier during construction; comparative validity rests on blinded human labels and frozen-packet Qwen rescoring.

Implementation. Contriever-v2 retrieves; `microsoft/deberta-v3-large` supplies the frozen coverage head and learned selector; `Llama-3.1-70B-Instruct` induces requirements and reads every packet; GPT-4o (`gpt-4o-2024-08-06`) verifies during construction; and Qwen2.5-72B only rescores frozen packets. Selector fitting, coverage calibration, and isotonic calibration use question-disjoint training subsets; all reported development examples and human cohorts are excluded. We set $(\lambda_1, \lambda_2, \lambda_3) = (0.01, 0.5, 0.2)$ and $(\tau_{\text{prop}}, \tau_{\text{qual}}, \tau_c) = (0.60, 0.58, 0.40)$, frozen before evaluation. Appendix A gives decoding settings, model roles, schemas, prompts, overflow handling, and hardware.

Metrics. Readers answer only from numbered passages and cite passage IDs. A deterministic span matcher aligns citations to benchmark support spans, with blinded adjudication for semantic or multi-passage matches. *Compiled-requirement coverage* is the fraction of automatically induced proposition, qualifier, and active-counterevidence records with a passing witness; it diagnoses compiler-conditioned construction, while majority-adjudicated human adequacy supplies the independent validity target. Harmful omission combines missing material qualifiers and available counterevidence. AUROC/AUPRC/ECE and risk–coverage use the same 600 adjudicated packets without population weighting; at 80% coverage each score approves its 480 highest-ranked packets. Answer F1 and citation precision/recall are downstream corroboration.

Uncertainty. Cells marked $\pm$ are five-seed means and standard deviations; human endpoints are fixed-packet point estimates with question- or packet-clustered intervals. Paired bootstrap intervals use 10,000 resamples, and the four prespecified PACKETCERT–CE+NLI contrasts receive Holm correction. The directly measured end-to-end timing aggregate contains 5000 queries at batch size and concurrency one.

Human protocols. The 600-packet cohort has 150 packets per dataset and three independent labels per packet. Separate non-overlapping 200-question HotpotQA and HotpotQA-Hard studies measure qualifiers ($\kappa = 0.74$) and omission categories ($\kappa = 0.71$). The near-tie study uses two first-pass judges and adjudicates disagreements. Judges see shuffled pool passages and anonymized packets, but no method, score, requirements, budget, or answer. Qualified QA/NLP annotators receive randomized order and $28/hour; Appendix A gives rubrics, cohorts, and adjudication.

5.2 RQ1–RQ2: CERTIFICATE DECISIONS AND MECHANISM ISOLATION

CE+NLI isolates the added value of relation-coupled construction because it shares the detector, activation rule, pool, reader, and cap. PACKETCERT raises compiled-requirement coverage by 11.1 points and citation precision by 4.1 points while reducing unconditional hidden counterevidence from 9.6% to 6.3%. More importantly, independent labels show an 8.9-point approval-precision gain. It also has the strongest adequacy endpoints and shortest realized packet among the fixed-pool methods in Table 1.

The prespecified near-tie analysis forms all $\binom{6}{2} \times 4$ datasets $\times 4$ caps $= 240$ pairs from Top-k, CE Rerank, CE+NLI, ECoRAG, CompSelect, and PACKETCERT. Before labeling, overlapping paired-bootstrap F1 intervals select 36 pairs. Their median F1 gap is 0.35 points (95% CI 0.22–0.47),

Table 1: Mechanism isolation with a shared top-50 pool; fixed-pool rows use a $B = 2.1$K maximum, while Iterative Retrieval is an adaptive reference. Compiled requirement coverage, HCE, abstention, and expansion use full HotpotQA; qualifier rates use the separate 200-question study; A-P/R uses the 600-packet cohort. GPT-4o Pass is a controller diagnostic.

(a) Direct packet endpoints (%)

Method	Comp. req. cov.	Qual.	HCE c/u	A-P/R	Abst./Exp.
CE Rerank	78.4±0.6	82.4	24.0±0.8/12.4	75.8/72.6	–
CE + NLI	81.7±0.5	86.9	18.6/9.6	82.5/77.9	–
ECoRAG	85.4±0.5	87.8	17.2/8.9	84.0/80.8	–
CompSelect	86.1±0.5	88.5	16.1/8.3	85.1/81.4	–
PACKETCERT	92.8±0.4	91.2	12.2/6.3	91.4/88.7	2.7/18.9

(b) Downstream quality and packet cost

Method	F1	Cite P/R (%)	HCE (%)	Pass (%)	Tok.; p95 ms
Top-k	71.2±0.5	52.3±0.8 / 78.1±0.6	18.7±0.7	71.4	2.1K; 24
CE Rerank	74.8±0.4	61.2±0.7 / 74.6±0.5	12.4±0.6	78.2±0.6	2.1K; 78
RECOMP	73.5±0.5	58.4±0.8 / 72.3±0.6	14.8±0.7	75.6	1.6K; 62
Iterative Retr.	76.3±0.4	63.1±0.6 / 79.4±0.5	10.8±0.5	80.1	2.4K; 198
CE + NLI	75.3±0.4	64.8±0.6 / 76.1±0.5	9.6±0.5	81.4±0.5	2.1K; 83
ECoRAG	76.5±0.4	65.9±0.6 / 78.6±0.5	8.9±0.5	83.1±0.5	1.6K; 102
CompSelect	76.8±0.4	66.4±0.6 / 79.1±0.5	8.3±0.5	83.7±0.5	1.5K; 96
PACKETCERT	77.4±0.4	68.9±0.5 / 81.2±0.4	6.3±0.4	85.7±0.4	1.4K; 84

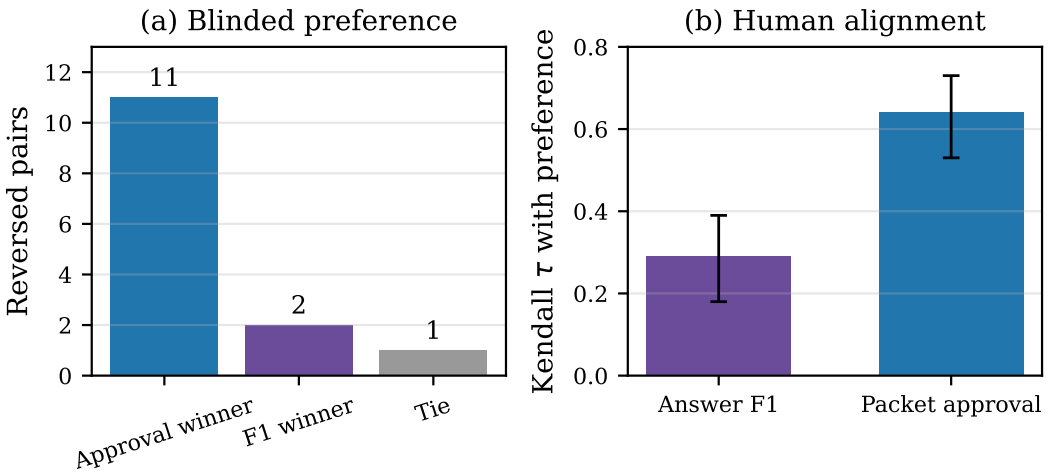


Figure 1: Conditional tie-breaking among answer-F1 near ties. A prespecified interval-overlap rule selects 36 of 240 dataset–budget–method-pair candidates; F1 and approval reverse 14. Blinded judges prefer the approval winner in 11 (78.6%; 95% Wilson CI 52.4–92.4), the F1 winner in 2, and tie once. Kendall-τ bars are paired-bootstrap 95% intervals.

versus 6.8 approval points (5.2–8.4). Across these conditional tie-break cases, approval aligns with preference at $\tau = 0.64$ (0.53–0.73), versus 0.29 (0.18–0.39) for F1.

5.3 RQ1: CERTIFICATE VALIDITY AND SELECTIVE RISK

The labels distinguish adequate, qualifier-missing, counterevidence-missing, and context-missing packets. The certificate score reaches AUROC 0.881 (95% CI 0.853–0.906) and cuts Risk@80 from 7.6% for SURE-RAG to 2.9%, a 4.7-point paired difference (2.6–6.9); Figure 3 reports the validated selective range and budget sweep.

Table 2: Harmful-omission detection on the fixed 600-packet, four-dataset cohort. Each packet has three labels before majority adjudication. Metrics are unweighted cohort estimates; Risk@80 gives harmful-omission counts among the 480 highest-ranked packets.

Score	AUROC	AUPRC	ECE ↓	Risk@80 ↓
Answer F1	0.662	0.621	0.151	15.2% [73/480]
Citation precision	0.731	0.698	0.119	13.1% [63/480]
RAGAS	0.758	0.724	0.105	11.4% [55/480]
ARES	0.779	0.746	0.092	10.2% [49/480]
RAGChecker	0.811	0.782	0.081	8.8% [42/480]
SURE-RAG	0.832	0.804	0.073	7.6% [36/480]
Certificate score	0.881	0.857	0.057	2.9% [14/480]

Table 3: Requirement fidelity and frozen-packet decision stability. The 240-question audit uses three blinded annotators; corrected requirements rescore identical pools and packets.

(a) Requirement-induction audit

Type	Slots	Precision	Recall	F1
Proposition	612	93.1	91.4	92.2
Qualifier	371	88.6	85.1	86.8
Counterevidence	184	85.9	82.6	84.2

(b) Corrected-requirement sensitivity

Requirements	Cite precision	HCE	Decision flips	Order stability
Original	68.9	6.3	–	–
Adjudicated	69.3	6.0	11/240	94.6%

(c) Verifier-family and prompt robustness

View	Pass rate	Human P/R	Agreement/κ	Flip/top-order
GPT-4o	85.7±0.4	91.4/88.7	–	–
Qwen2.5-72B	84.9 ± 0.5	90.6/87.9	86.7%/0.72	13.3%
Three prompt variants	84.2–86.1	90.1/87.4	0.69–0.74	14/15

In the separate HotpotQA study, PACKETCERT preserves qualifiers in 91.2% (547/600; 95% CI 88.7–93.2), versus 82.4% (494/600) for CE Rerank and 76.8% (461/600) for Top-k. Harmful omission is 3.1% (19/600; 2.0–4.9) versus 8.7% (52/600; 6.5–11.2) for CE; rates and fractions use the same 600 judgments.

5.4 RQ3: REQUIREMENT, VERIFIER, AND CLUTTER STABILITY

Table 3 shows that most requirement-induction errors merge nested qualifiers or miss implicit temporal scope; adjudicated requirements move citation precision by only 0.4 points and unconditional omission by -0.3 points. Paired-bootstrap method order is preserved in 94.6% of replicates. Frozen-packet Qwen agreement is 86.7% with $\kappa = 0.72$, and PACKETCERT leads the diagnostic approval ordering in 14 of 15 family/prompt/seed aggregates.

Under controlled clutter, PACKETCERT improves citation precision by 5.8 points and lowers hidden-counterevidence omission by 6.0 points relative to CE+NLI, while achieving 67.9 F1 at 1.6K tokens. Figure 2 uses the separately annotated 200-question subset; Table 8 uses the complete HotpotQA-Hard split, which explains the 7.4% subset and 9.8% full-split values.

5.5 RQ4: NATURAL LONG-FORM TRANSFER

On ASQA, PACKETCERT improves over CompSelect by 0.7 Disambig-F1 (95% CI 0.3–1.1; Holm-adjusted $p = 0.008$) and 1.1 omission points (CI 0.4–1.8; $p < 0.001$). Long-Context gains 0.7 F1

Table 4: ASQA validation on all 948 examples. Panel (a) reports five-seed mean ± standard deviation. Panel (b) reports three blinded omission judgments for one prespecified frozen packet per method–example pair, with example-clustered inference, plus evidence tokens and pooled p95 milliseconds.

(a) Answer and citation quality

Method	Disambig-F1	ROUGE-L	Cite P/R	Qual.
CE Rerank	31.6±0.5	33.8±0.4	55.2±0.7/62.8±0.6	81.6±0.6
CE + NLI	32.0±0.4	34.2±0.4	61.7±0.6/66.5±0.6	86.7±0.5
ECoRAG	32.7±0.4	35.1±0.3	63.8±0.6/69.6±0.5	89.1±0.5
CompSelect	32.6±0.4	35.0±0.3	64.4±0.6/69.1±0.5	89.5±0.4
PACKETCERT	33.3±0.3	35.8±0.3	67.1±0.5/72.0±0.5	93.4±0.4
Long-Context	34.0±0.3	36.2±0.3	58.6±0.6/73.1±0.4	88.6±0.5

(b) Omission and packet cost

Method	Harmful omission	Tokens	Packet p95	Omission judgments
CE Rerank	13.2%	2.1K	79	376/2844
CE + NLI	9.4%	2.1K	84	267/2844
ECoRAG	7.5%	1.7K	107	213/2844
CompSelect	7.0%	1.5K	99	199/2844
PACKETCERT	5.9%	1.6K	89	168/2844
Long-Context	8.1%	4.2K	55	230/2844

Table 5: Measured operating points. D/S/V/E/G denote decomposition, selection, gating, mean expansions, and generation. "Pkt" is local scoring/packing; E2E includes every remote call. Gen. In/Out is the reader call, whereas USD/query is cumulative measured billing. Approval is the GPT-4o controller diagnostic; blinded judgments provide comparative validity.

Method	Stages D/S/V/E/G	Pkt/E2E p50/p95	Gen. In/Out; USD	Appr./Exp./Abst.
CE Rerank	0/1/0/0/1	44/78; 1180/1910	2.1K/67; 0.031	78.2/–/–
CE + NLI	0/1/1/0/1	48/83; 1310/2130	2.1K/66; 0.036	81.4/–/–
ECoRAG	0/1/1/0.23/1	58/102; 1440/2360	1.7K/68; 0.034	83.1/23.0/1.9
CompSelect	0/1/1/0/1	53/96; 1390/2250	1.5K/66; 0.033	83.7/–/–
Long-Context	0/0/0/0/1	31/55; 1730/2880	4.2K/71; 0.052	88.9/–/–
PACKETCERT	1/1/1/0.19/1	49/84; 1490/2410	1.6K/69; 0.038	85.7/18.9/2.7

with 2.6K additional evidence tokens, while PACKETCERT gains 8.5 citation-precision points and lowers omission by 2.2 points.

5.6 RQ2–RQ4: CONSTRUCTION COST AND COMPONENT ROLES

PACKETCERT's full path has 2.41 s p95, 0.28 s above CE+NLI. Its reader call uses 1.6K input tokens, including 1.4K evidence tokens, while cumulative measured billing is 26.9% below Long-Context. The 49/84 ms column isolates local coverage scoring and packing; E2E includes induction, verification, expansion, and generation.

The ablation identifies distinct component roles. Removing the anti-omission term raises hidden counterevidence from 6.3% to 15.1% with only 0.3 F1 loss; removing qualifier slots reduces scoped citation support; removing the ILP teacher weakens the selector prior and causes the largest F1 drop. A same-family verifier gives the weakest citation/omission pair among verifier variants.

6 DISCUSSION

Three findings connect mechanism to consequence. The shared-safeguard control shows that retaining NLI-flagged passages independently does not recover joint evidence coverage. Blinded adequacy labels validate the certificate score without exposing requirements or verifier outputs, and the conditional near-tie study shows where that score changes a decision that F1 cannot resolve. Corrected

requirements, Qwen rescoring, prompt variants, clutter, four caps, and ASQA preserve the principal adequacy ordering.

The evidence covers fixed candidate pools, four QA workloads, and the tested induction and verifier families. Learned semantic judgments remain hardest for implicit temporal scope and unrecognized counterevidence, while retrieval determines which facts are available to any constructor. Within this regime, UNCERTIFIEDUNDERBUDGET, witness IDs, and margins make failed coverage explicit, and blinded labels show that packet approval detects harmful omission substantially better than answer F1.

7 CONCLUSION

PACKETCERT makes compact selection preserve the set-level evidence that generation inherits: linked propositions, governing qualifiers, and available counterevidence. Independent four-dataset labels yield 91.4/88.7% approval precision/recall and 2.9% Risk@80; HotpotQA compiled-requirement coverage reaches 92.8% at 1.4K tokens. Approval also wins 11 of 14 blinded preference judgments when it reverses an answer-F1 near tie, and ASQA improves answer quality and omission together. Set-level certification is therefore a practical construction objective and tie-break criterion for fixed-pool compact RAG.

ETHICS STATEMENT

Adult annotators evaluated public QA text with consent, withdrawal rights, prespecified compensation, and no personal or sensitive information; results are aggregate and method identities were hidden.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the objective, algorithm, controls, metrics, and uncertainty; Appendices A–F specify interfaces, cohorts, pseudocode, and reported coordinates.

AI USE STATEMENT

Generative AI tools assisted manuscript preparation, including drafting/editing, formatting, and bibliography checking. The authors reviewed the manuscript and take responsibility for its contents.

REFERENCES

Vašek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233. URL https://doi.org/10.1287/moor.4.3.233.

Davide Di Gioia. Entropic claim resolution: Uncertainty-driven evidence selection for RAG. *arXiv preprint arXiv:2603.28444*, 2026. URL https://arxiv.org/abs/2603.28444.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2587–2601. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.187. URL https://aclanthology.org/2022.naacl-main.187/.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998. doi: 10.1145/285055.285059. URL https://doi.org/10.1145/285055.285059.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Yeonseok Jeong, Jinsu Kim, Dohyeon Lee, and Seung-won Hwang. ECoRAG: Evidentiality-guided compression for long context RAG. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26607–26628. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1365. URL https://aclanthology.org/2025.findings-acl.1365/.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval-augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Vladimir Karpukhin, Barlas Oguz, 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*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. Association for Computing Machinery, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Aiyu Li, Qian Peng, and Bin Chen. Efficiency vs. verifiability in evidence-aware RAG: Does prompt compression preserve citation grounding? In *Proceedings of the Second Workshop on Customizable NLP*, pp. 202–215. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.customnlp4u-1.19. URL https://aclanthology.org/2026.customnlp4u-1.19/.

Jacob Menick, Maja Trębacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. URL https://arxiv.org/abs/2203.11147.

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. *Proceedings of the 2023 Conference on Empirical Methods*

in Natural Language Processing, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. doi: 10.48550/arXiv.1901.04085. URL https://arxiv.org/abs/1901.04085.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026. URL https://arxiv.org/abs/2605.03534.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 21999–22027, 2024. doi: 10.52202/079017-0692. URL https://doi.org/10.52202/079017-0692.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338–354. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. In *Proceedings of the ACM Web Conference 2026*, pp. 1971–1982. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3792158. URL https://doi.org/10.1145/3774904.3792158.

A OPERATIONAL INTERFACES AND EVALUATION VALIDITY

A.1 MODEL ROLES AND DATA SEPARATION

Table 6 separates construction, certification, and evaluation. The inducer and reader share a Llama checkpoint but never approve packets; GPT-4o has a separate parameter and prompt path from DeBERTa; Qwen only rescores frozen packets. GPT-4o enters PACKETCERT's controller and scores baselines post hoc, so its pass rate diagnoses the controller; blinded judgments and frozen-packet Qwen results carry comparative validity.

Table 6: Model roles and visible inputs. Gold answers and generated answers are unavailable to every pre-generation construction stage.

Component	Checkpoint / optimization	Inputs and outputs
Pool retriever	Contriever-v2, frozen	Query $\rightarrow$ shared top-50 passages
Requirement inducer	Llama-3.1-70B-Instruct, $T = 0$	Query $\rightarrow$ linked proposition and qualifier records
Coverage head	DeBERTa-v3-large; fitted then frozen	Query, linked requirement, passage $\rightarrow v_{ij}$
Selector head	DeBERTa-v3-large; Eq. 4	Query, all linked requirements, passage, token cost $\rightarrow s_i, \widehat{z}_i$
Typed-cover teacher	Gurobi, training only	Thresholded coverage graph and costs $\rightarrow$ exact membership labels
Set verifier	GPT-4o-2024-08-06, $T = 0$	Query, linked requirements, unordered packet $\rightarrow \nu, V, D_V$
Alternate verifier	Qwen2.5-72B-Instruct, $T = 0$	Same frozen verifier input; robustness analysis only
Reader	Llama-3.1-70B-Instruct	Query and numbered packet $\rightarrow$ cited answer

Training examples come only from official training splits. Question IDs separate coverage fitting, selector optimization, and threshold/isotonic calibration; no reported development example enters them. The 600-packet cohort and 240-question correction audit are sampled only after freezing parameters. Coverage fitting uses training support-witness pairs for each serialized linked record; that encoder/head is frozen before edges, exact teacher sets, or selector targets are computed. Equation 4 then updates only the scalar selector head from fixed v_{ij} and exact membership labels. At inference, $\widehat{z}_i$ modulates Eq. 5, while the separately evaluated v_{ij} controls hard certificate validity.

A.2 REQUIREMENT AND VERIFIER RECORDS

The invariant induction instruction is answer-free and asks the model to decompose the query into atomic propositions, attach every material temporal, entity, quantity, or comparison qualifier to its governing proposition ID, and avoid answering the query. The validated output is an ordered JSON-equivalent record

$$\{\texttt{propositions} : [(\texttt{id}, \texttt{text})], \quad \texttt{qualifiers} : [(\texttt{id}, \texttt{parent_id}, \texttt{text})]\}.$$

Grounding serializes each proposition alone and each qualifier together with its parent statement before scoring it against the fixed pool. NLI discovery then adds $\texttt{counter} = (\texttt{id}, \texttt{target_id}, \texttt{passage_id}, \texttt{text})$ when a pool passage contests a grounded proposition or qualifier above η_{act}. Stable IDs are preserved through the loss, greedy rule, verifier diagnostics, and returned certificate. Missing required fields invalidate the record rather than silently dropping an obligation.

The invariant verifier instruction asks whether the unordered passages jointly establish every proposition, preserve each attached qualifier, and expose every listed counterevidence item, without answering the query. Its validated output contains only

$$\{\texttt{score} : \nu, \quad \texttt{verdict} : V, \quad \texttt{deficient_ids} : D_V\}.$$

The three robustness prompts retain this schema and decision criteria while changing the inspection order: proposition-first, qualifier-first, or counterevidence-first. A schema failure cannot yield approval; the controller treats it as rejection and targets the lowest normalized-margin requirement when no valid deficient ID is available.

Schematic certificate. For a two-hop query asking which city hosted an event in a specified year, the compiler can emit $r_1 =$ "entity x held event e," $r_2 =$ "e occurred in city y," and qualifier $r_3 =$ "the governing edition is year t," with parent r_2. If the pool also contains a passage for a different edition, it activates counter slot r_4 targeting r_3. A packet with witnesses for r_1 and r_2 still fails when it omits r_3 or r_4; a passage mentioning the year without its parent relation does not satisfy r_3. This example illustrates why passage relevance and an independent contradiction flag do not implement Eq. 2.

A.3 Human targets and cohort accounting

Judges label the text that a reader would receive, not the induced requirement graph. A packet is *adequate* when its passages jointly support a complete answer, preserve every material scope-changing qualifier visible in the common pool, and expose pool evidence that would materially change or contest that answer. Otherwise the primary failure label is qualifier-missing, counterevidence-missing, or context-missing. Harmful omission combines the first two failure types. Pairwise judges apply the same criteria comparatively and may return a tie; automatic requirements, verifier outputs, method identity, budget, and generated answers remain hidden.

Table 7: Human-evidence cohorts and inferential units. Counts refer to sampled packets or questions; annotator votes are clustered before interval estimation.

Study	Sample and labels	Sampling, decision, and uncertainty
Adequacy / selective risk	600 packets; three independent labels each	150 per dataset; stratified by method, budget, and score decile; majority adjudication; unweighted cohort estimates; packet bootstrap
HotpotQA qualifiers	200 questions; three labels per method	Stratified-random original split; empirical rates and question-clustered intervals
HotpotQA-Hard omissions	200 questions; two evaluators	Controlled-clutter subset; blinded labels; $\kappa = 0.71$
Operating-point preference	36 pairs; two first-pass judges plus adjudicator	Deterministic filter from 240 candidates before labeling; pair-level ordinal outcomes and paired-question bootstrap

For the conditional near-tie cohort, all unordered method pairs share a dataset and cap. The deterministic F1 interval-overlap rule selects 36 of 240 candidates before labeling; 14 reverse under approval. Judges prefer the approval winner in 11 reversals, the F1 winner in two, and tie once (78.6% approval-win rate; 95% Wilson CI 52.4–92.4). This estimand concerns F1-indistinguishable choices rather than unrestricted model ranking.

B Controller pseudocode and certificate-evidence map

The matrix maps each protected object to its certificate obligation, enforcement boundary, and direct evidence. The stressor may influence construction, while the separately parameterized verifier remains outside the selector's parameter and prompt path.

Certificate-evidence map; "pool fixed" is the shared trust boundary.

Algorithm 1 PACKETCERT: synthesize and verify a typed evidence certificate

Require: Query q, fixed pool $\mathcal{P}_q$, budget B
$\mathcal{R}_q \leftarrow \mathrm{InduceAndGround}(q, \mathcal{P}_q)$
$v_{ij} \leftarrow \sigma(\mathrm{Coverage}(p_i, r_j))$, $\widehat{z}_i \leftarrow \sigma(\mathrm{Selector}(q, \mathcal{R}_q, p_i, |p_i|_{\mathrm{tok}}))$
$\mathcal{S} \leftarrow \emptyset$; $D_V \leftarrow \emptyset$
loop
  $U \leftarrow \{r_j : c_j(\mathcal{S}) < \tau_{t(j)}\}$
  if $U = \emptyset$ **then**
    $(\nu, D_V) \leftarrow \mathrm{Verify}(\mathcal{S}, q, \mathcal{R}_q)$
    if $\nu \geq \tau_V$ **then**
      return $(\mathcal{C}_B(q), \textsc{Certified})$
    end if
    if $D_V = \emptyset$ **then**
      $D_V \leftarrow \{\arg\min_j c_j(\mathcal{S})/\tau_{t(j)}\}$
    end if
  end if
  $F \leftarrow \{p_i \notin \mathcal{S} : T(\mathcal{S}) + |p_i|_{\mathrm{tok}} \leq B\}$
  if $F = \emptyset$ or Eq. 5 has zero gain **then**
    return $(\mathcal{S}, \textsc{UncertifiedUnderBudget})$
  end if
  $p^* \leftarrow$ Eq. 5; $\mathcal{S} \leftarrow \mathcal{S} \cup \{p^*\}$
  $D_V \leftarrow \emptyset$
end loop

Failure surface	Certificate obligation	Boundary, strongest comparison, setting, cost, and evidence
Dropped qualifier	Material-qualifier coverage	Pool fixed; CompSelect; corrected-slot audit; added tokens; residual implicit-scope errors; Eq. 2, Table 3.
Counterevidence suppression	Available counterevidence exposed	Pool fixed and NLI shared; CE+NLI; controlled clutter; verification overhead; residual NLI misses; Tables 1, 8.
Partial multi-hop packet	Joint proposition coverage	Pool fixed; ECoRAG; four datasets; compact token use; residual decomposition merges; Table 1, Figure 3.
Self-approval	Separate-parameter set sufficiency	Selector/verifier families separated; same-family variant; frozen-packet Qwen rescore; API latency; prompt flips; Table 3.
Clutter and distractors	Stable approve/reject ordering	Pool fixed; safeguarded reranker; HotpotQA-Hard; measured packet latency; residual lexical aliases; Figure 2.
Budget exhaustion	Explicit abstention	Hard token cap; all builders; 0.8–2.0K sweep; coverage tradeoff; residual low-recall pools; Figure 3, Table 5.

C CERTIFICATION INVARIANT AND TEACHER RELATION

Proposition (sound return). If Algorithm 1 returns CERTIFIED, its packet satisfies every conjunct of Eq. 2.

Proof. Verification is called only when the hard uncovered set is empty, so each typed requirement has a single witness meeting its type threshold. Every addition is drawn from $F(\mathcal{S})$, preserving $T(\mathcal{S}) \leq B$. The CERTIFIED branch additionally requires $\nu \geq \tau_V$. These are exactly the approval conjuncts. □

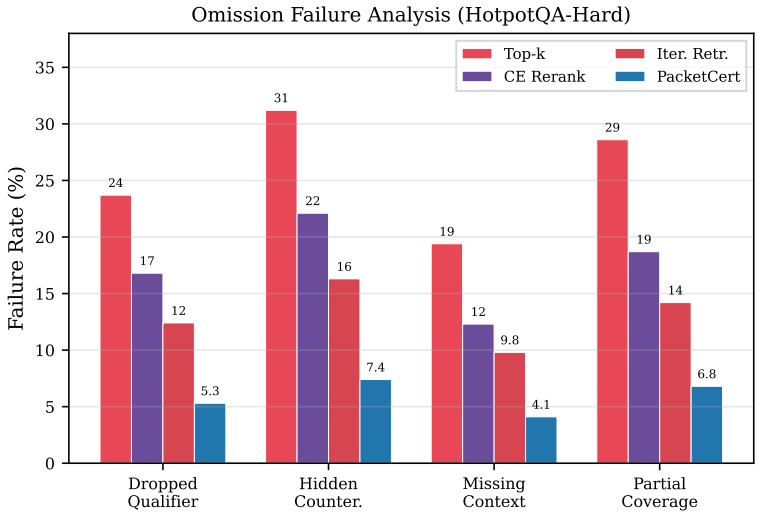


Figure 2: Two-evaluator audit of 200 HotpotQA-Hard questions, distinct from the three-annotator random-HotpotQA study. PACKETCERT lowers every omission category; hidden counterevidence is 7.4% versus 22.1% for CE Rerank on this subset.

The exact ILP teacher minimizes cost for the frozen thresholded training graph. The familiar H_m bound characterizes unbudgeted weighted greedy cover without $\widehat{z}_i$ modulation (Chvátal, 1979). Algorithm 1 instead guarantees the sound-return proposition above while its remaining-budget filter and verifier-directed expansion produce an explicit uncertified outcome when certification is unavailable.

Table 8: HotpotQA-Hard controlled clutter stress test ($K = 80$). CE+NLI shares PACKETCERT's safeguard. F1, citation precision, and HCE are five-seed means $\pm$ standard deviation; Pass is the pooled GPT-4o controller diagnostic.

Method	F1	Cite Prec.	HCE ↓	Pass	Tokens
Top-k	58.3±0.8	41.2±1.1	28.4±0.9	58.7	2.1K
CE Rerank	63.7±0.7	49.8±0.9	21.6±0.8	66.4	2.1K
CE + NLI	64.5 ± 0.7	55.6±0.8	15.8±0.7	72.8 ± 0.6	2.1K
RECOMP	61.4±0.8	46.3±1.0	24.1±0.9	63.1	1.6K
ECoRAG	65.4 ± 0.6	58.1±0.7	13.2±0.6	75.1 ± 0.6	1.7K
CompSelect	65.8 ± 0.6	58.7±0.7	12.6±0.6	75.9 ± 0.5	1.6K
Iterative Retr.	66.2±0.6	53.7±0.8	17.9±0.7	71.3	2.7K
Long-Context	68.8±0.5	47.1±0.7	16.2±0.6	69.8	4.3K
PACKETCERT	67.9±0.5	61.4±0.6	9.8±0.5	79.2	1.6K

The full-split clutter result complements the 200-question failure taxonomy above. PACKETCERT's 61.4 citation precision and 9.8% hidden-counterevidence omission improve jointly over every compact baseline, while its 67.9 F1 remains within 0.9 points of the 4.3K-token Long-Context reference at 1.6K tokens.

Table 9: HotpotQA ablation. Removing the anti-omission penalty raises hidden counterevidence from 6.3% to 15.1% with only 0.3 F1 loss.

Variant	F1	Cite Prec.	HCE ↓	Tokens
PACKETCERT	77.4	68.9	6.3	1.4K
– Anti-omission penalty	77.1	66.2	15.1	1.2K
– Qualifier slots	76.8	65.4	7.8	1.3K
– ILP teacher	76.2	66.8	7.1	1.5K
– Independent verifier	77.3	68.1	6.8	1.4K
Same-family verifier	76.9	64.3	9.2	1.4K

The ablation shows different endpoint sensitivities across answer utility and evidence failure. The anti-omission term has the largest movement in hidden counterevidence, qualifier-slot removal most directly changes scoped citation support, and teacher removal produces the largest F1 decrease. The verifier rows compare independent and same-family final set checks.

D SELECTIVE-RISK AND BUDGET COORDINATES

The selective analysis holds the adjudicated packet cohort fixed and varies only the score threshold; the budget analysis varies the constructor cap on the same four-dataset evaluation design.

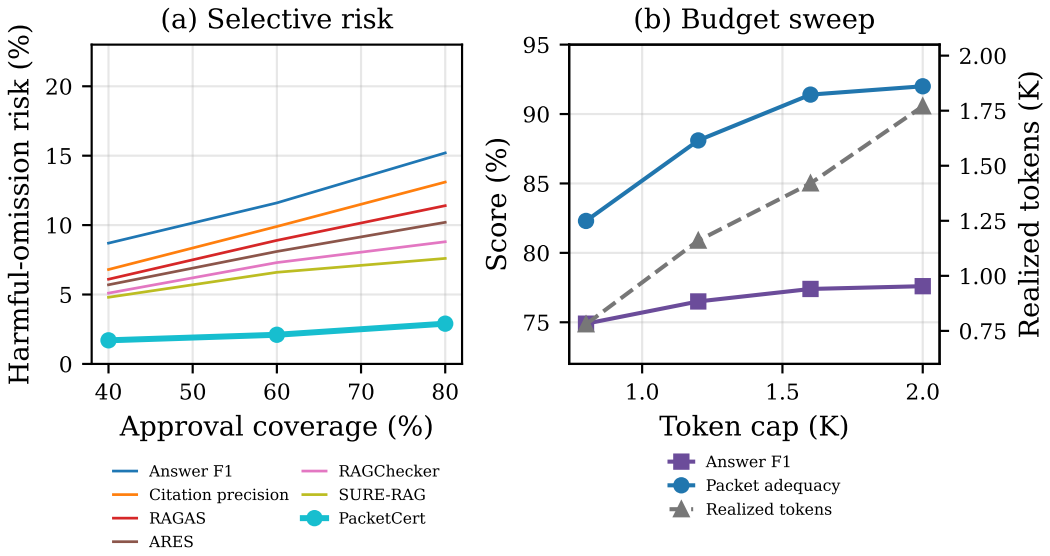


Figure 3: Certificate risk versus coverage and token budget. (a) On the validated 40–80% selective range, the certificate score has the lowest harmful-omission risk. (b) adequacy rises through 1.6K tokens after F1 begins to saturate; realized length remains below each cap.

Across the reported 40–80% selective range, the certificate score retains the lowest harmful-omission risk. The budget coordinates show a complementary saturation pattern: answer F1 changes by 0.2 points from 1.6K to 2.0K, whereas packet adequacy still rises by 0.6 points.

E HUMAN-LABEL COUNTS AND ADJUDICATION

The risk cohort contains 150 packets from each dataset and is stratified by method, budget, and score decile. Three labels per packet are majority-adjudicated; all reported classification and selective-risk

Table 10: Harmful-omission risk (%) for Figure 3(a). Every score ranks the same 600 adjudicated packets; results are finite-cohort, unweighted estimates.

Score	40% cov.	60% cov.	80% cov.
Answer F1	8.7	11.6	15.2
Citation precision	6.8	9.9	13.1
RAGAS	6.1	8.9	11.4
ARES	5.7	8.1	10.2
RAGChecker	5.1	7.3	8.8
SURE-RAG	4.8	6.6	7.6
Certificate score	1.7	2.1	2.9

Table 11: PACKETCERT budget coordinates for Figure 3(b).

Hard cap	0.8K	1.2K	1.6K	2.0K
Answer F1	74.9	76.5	77.4	77.6
Packet adequacy	82.3	88.1	91.4	92.0
Realized tokens	0.78K	1.16K	1.42K	1.77K

values are unweighted estimates for this fixed cohort. "Adequate" is positive, and harmful omission combines qualifier-missing and counterevidence-missing outcomes. At 80% coverage, each score contributes exactly 480 approved packets. The HotpotQA qualifier study has three judgments for each of 200 questions; rates and parenthetical counts use the same 600 judgments, with question-clustered intervals. CE qualifier preservation is 79.1–85.2 and Top-k is 73.3–80.0. The two 200-question studies do not overlap.

F CONTEXTUAL ANSWER-QUALITY VIEW

Table 12: Contextual multi-hop answer quality. Each cell is answer F1 / average input tokens, providing downstream context alongside the direct certificate endpoints.

Method	HotpotQA	MuSiQue	2Wiki	Average
Top-k	71.2±0.5 / 2.1K	48.3±0.7 / 2.1K	62.7±0.4 / 2.1K	60.7 / 2.1K
MMR	72.4±0.4 / 2.0K	49.8±0.6 / 2.0K	63.9±0.4 / 2.0K	62.0 / 2.0K
CE Rerank	74.8±0.4 / 2.1K	52.4±0.6 / 2.1K	66.1±0.4 / 2.1K	64.4 / 2.1K
RECOMP	73.5±0.5 / 1.6K	50.7±0.7 / 1.5K	64.8±0.5 / 1.5K	63.0 / 1.5K
LLMLingua	72.1±0.5 / 1.3K	48.9±0.7 / 1.2K	63.2±0.5 / 1.3K	61.4 / 1.3K
Iterative Retr.	76.3±0.4 / 2.4K	54.1±0.6 / 2.6K	67.8±0.4 / 2.5K	66.1 / 2.5K
Long-Context	78.3±0.3 / 4.1K	55.7±0.5 / 4.2K	69.4±0.3 / 4.0K	67.8 / 4.1K
PACKETCERT	77.4±0.4 / 1.4K	54.8±0.5 / 1.5K	68.7±0.3 / 1.3K	67.0 / 1.4K

The multi-hop view places the direct certificate results in downstream context. PACKETCERT averages 67.0 F1 at 1.4K tokens across HotpotQA, MuSiQue, and 2WikiMultiHopQA, within 0.8 points of Long-Context while using roughly one third of its input.

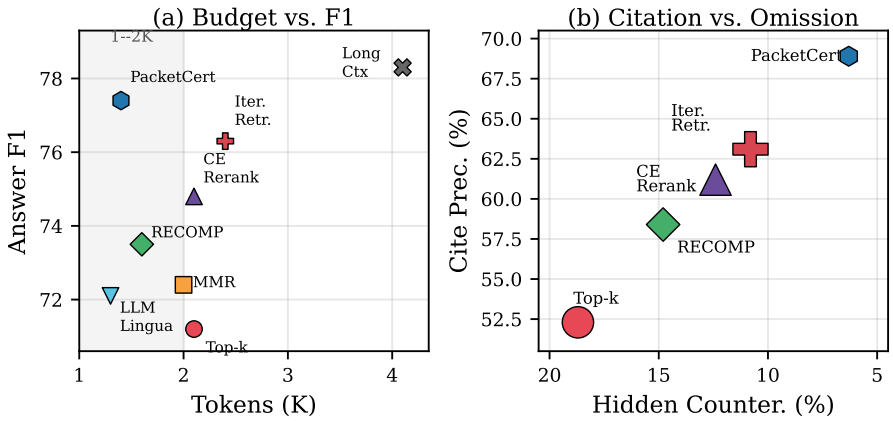


Figure 4: Contextual HotpotQA operating points. Left: answer F1 versus average input tokens. Right: citation precision versus hidden counterevidence for Top-k, CE Rerank, RECOMP, Iterative Retrieval, and PACKETCERT, using the values in Table 1 without interpolation.

Panel (a) separates compactness from answer quality: PACKETCERT reaches 77.4 F1 at 1.4K tokens, within 0.9 points of the 4.1K-token Long-Context reference and above every other displayed compact method. Panel (b) orients lower hidden counterevidence to the right, so the desirable region is the upper-right corner. PACKETCERT occupies that region with 68.9% citation precision and 6.3% hidden counterevidence, jointly improving both axes over the displayed controls.

These coordinates explain why answer F1 alone is insufficient for compact-packet selection. The closest answer-quality point can require substantially more context, while similarly compact points can conceal evidence that changes the supported answer. The certificate endpoints expose this second dimension directly.