UNITALIGN: CITATION REPAIR THAT PRESERVES SUPPORTED CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A citation-grounded answer can contain one unsupported sentence amid several that already have valid evidence. Regenerating the answer reopens those verified claims, whereas citation scoring alone leaves the failed claim untouched. UNITALIGN turns the sentence and its cited spans into a shared interface for human inspection, verifier decisions, and corrective action. A budgeted packet compiler removes redundant evidence; sentence-level provenance then permits a calibrated external verifier to trigger focused retrieval and rewrite only the failed unit. On BioMed-180, nine Llama-3-70B-Instruct systems receive the same corpus, citation schema, and 8,192-token visible-evidence allowance. In 360-pair expert audits per system, UNITALIGN reaches 94.7% supported-sentence rate and 93.1% citation precision—gains of 7.6 and 11.7 points over matched-stack RARR—and exceeds the strongest faithful native citation system by 2.4 and 1.7 points. A fixed-state intervention sharing initial answers, citation bindings, verifier decisions, and triggers shows why: the local policy improves support on triggered units by +6.1 [+1.4, +10.9] points, preserves +44.2 [+38.1, +50.5] more previously supported content, and saves 7.01 [6.42, 7.63] s relative to whole-answer regeneration. Failure locality—unsupported-unit density, concentration, and discourse coupling—also reveals the escalation boundary: local repair leads by 8.9 support points for sparse, weakly coupled failures, while answer-wide repair leads by 5.8 for dense, strongly coupled failures. Complete public audits and Qwen and Mixtral generators preserve the comparative advantage; verifier swaps retain 93.2–95.2% human SSR. Fine-grained attribution can therefore govern corrective action, not merely document it.

1 INTRODUCTION

Consider a cited medical answer with three claims: the contraindication threshold is attached to a passage about dose adjustment, while the surrounding monitoring and later-line treatment sentences are correctly supported. The answer is neither wholly reliable nor wholly defective. A repair system must correct the first sentence without needlessly changing the other two. This mixed-validity case is common to long-form attributed generation, yet answer-level regeneration treats every sentence as mutable and citation-only scoring never changes the unsupported claim.

The missing interface is a shared unit for inspection and intervention. Existing systems establish rich provenance surfaces: ALCE produces citation-bearing long-form answers (Gao et al., 2023b); ReClaim and verifiable generation bind evidence to sentences and subsentences (Xia et al., 2025; Cao & Wang, 2024); and CAGE maps answer units to support subgraphs (Yan et al., 2026). CiteFix, VeriCite, and FineRef improve citation assignment, verification, or correction (Maheshwari et al., 2025; Qian et al., 2025; Peng et al., 2026), while RARR and Self-RAG revise generation through retrieval and reflection (Gao et al., 2023a; Asai et al., 2024). What remains unresolved is how a post-generation support decision should delimit the content that can change. That choice jointly determines reliability, preservation of valid text, and repair cost.

UNITALIGN resolves this interface mismatch through *audit–action alignment*: the delivered sentence and its cited evidence are simultaneously the object shown to human auditors, the input to an external verifier, and the minimum mutable unit. Figure 1 shows the resulting control path. A coverage-aware packet compiler fits diverse evidence under a fixed visible budget; the generator exposes sentence-level provenance; and a calibrated verifier sends only unsupported units through focused retrieval and

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.

constrained rewriting. Passed sentences remain fixed. The method therefore converts citation support from an offline score into a selective control signal while preserving the exact answer surface that readers already inspect.

The evaluation establishes both the comparative gain and its mechanism. On BioMed-180, nine systems share Llama-3-70B-Instruct, the corpus, the sentence-final citation schema, and an 8,192-token visible-evidence allowance. UNITALIGN improves expert-audited SSR from 87.1% to 94.7% and citation precision from 81.4% to 93.1% over matched-stack RARR, with a paired helpfulness gain of +0.27 [+0.16, +0.38]. It also leads the strongest faithful native citation system by 2.4 SSR and 1.7 citation-precision points. In a fixed-state intervention over 180 queries and 987 initial units, the local and whole-answer branches share the initial answer, citation bindings, verifier scores, and 82 triggers. Local repair adds +6.1 [+1.4, +10.9] support points on triggered units, preserves +44.2 [+38.1, +50.5] more passed content, and saves 7.01 [6.42, 7.63] s end to end. These are complete-policy effects: branch-specific retrieval and generation work are measured as outcomes rather than treated as matched inputs. Across BioMed-180 and HotpotQA-Citation, failure locality—unsupported-unit density, contiguous-run concentration, and discourse coupling—explains when that advantage reverses: local repair leads by 8.9 points in the sparse, weak-coupling cell, whereas answer-wide repair leads by 5.8 in the dense, strong-coupling cell.

The paper contributes (i) an aligned sentence–evidence control unit that connects human audit, calibrated verification, and selective rewriting; (ii) a budgeted evidence compiler and local repair loop that operationalize this unit; (iii) failure locality and a fixed-state intervention that resolve the local-to-global repair crossover through support, preservation, edit-spillover, and latency outcomes; and (iv) expert and holistic validation on BioMed-180, three complete public citation benchmarks, faithful native systems, alternative verifiers, and Llama, Qwen, and Mixtral generators. Together, these results show that fine-grained provenance can govern corrective action, not merely document it.

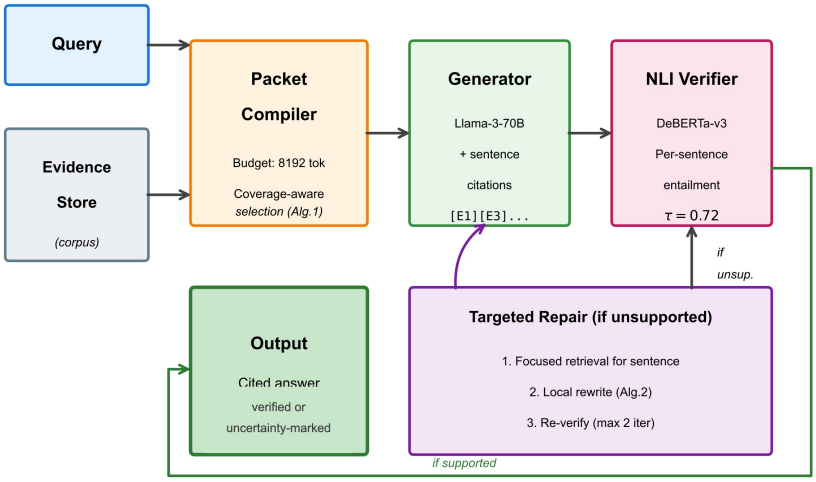


Figure 1: UNITALIGN aligns three boundaries. The packet compiler selects evidence under a fixed budget, the generator emits sentence-level provenance, and the verifier evaluates the same sentence–evidence unit shown to auditors. Failed units trigger focused retrieval and rewriting; passed units are copied unchanged. Unresolved units remain explicitly uncertainty-marked.

2 RELATED WORK

Retrieval-augmented generation supplies evidence to a generator (Lewis et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2021), and ALCE established systematic evaluation for citation-bearing long-form answers (Gao et al., 2023b). More recent systems expose finer provenance before an answer is delivered: ReClaim alternates reference and claim generation at sentence granularity (Xia et al., 2025), verifiable generation constructs subsentence citations (Cao & Wang, 2024), and

CAGE realizes answer units from support subgraphs (Yan et al., 2026). UNITALIGN starts after this provenance surface exists. Its technical question is whether the exposed unit can also serve as the boundary of corrective action.

Verification and citation correction provide the closest components. Entailment models score support for scientific claims (Wadden et al., 2020; 2022b;a), and post-hoc factuality metrics identify unsupported long-form content (Min et al., 2023; Manakul et al., 2023). CiteFix reassigns citations without rewriting claims (Maheshwari et al., 2025); VeriCite verifies evidence before final-answer refinement (Qian et al., 2025); and FineRef learns fine-grained reflection and correction (Peng et al., 2026). UNITALIGN combines these otherwise separate roles in a specific contract: a calibrated external decision on the delivered sentence–evidence pair triggers retrieval and rewriting for that pair, while every passed pair is immutable. Accordingly, the native-system experiment compares final support under published generation contracts, whereas the fixed-state experiment isolates the complete downstream consequences of local and answer-wide repair policies.

Active retrieval, Self-RAG, RARR, and reflection loops adapt generation after intermediate feedback (Jiang et al., 2023; Asai et al., 2024; Gao et al., 2023a; Shinn et al., 2023). Their controllers may retrieve, critique, or revise at several points in generation. The distinction is the control object: ALCE and ReClaim construct provenance before delivery; CiteFix changes citation links without claim text; VeriCite synthesizes an answer after evidence verification; FineRef trains correction from reflected errors; and RARR revises an answer through research and revision. UNITALIGN instead keeps the delivered sentence–evidence pair stable across annotation, verification, and rewriting, with every passed pair frozen.

Together, these lines of work supply provenance, support signals, and revision mechanisms. UNITALIGN contributes the interface that joins them: one externally auditable unit determines what is checked, what triggers retrieval, what can be rewritten, and what must be preserved. Failure locality then characterizes when correction can remain at that interface and when cross-sentence dependence favors answer-wide revision. This definition supports two complementary evaluations without conflating their contracts—faithful native comparisons for generation-time grounding and a fixed-state local-versus-answer-wide intervention for post-generation repair.

3 UNITALIGN: ALIGNED VERIFICATION AND REPAIR

3.1 OBJECTIVE AND ALIGNED DECISION UNIT

Given query q and corpus $\mathcal{D}$, UNITALIGN seeks a supported cited answer under visible-evidence allowance B while preserving passed text. It segments answer A into sentences s_i carrying cited spans $\{e_{ij}\}$ and defines $u_i = (s_i, \{e_{ij}\})$. Human auditors inspect u_i, the verifier accepts it when $\mathcal{V}(u_i) \geq \tau$, and repair may replace only s_i and its citations. The packet compiler controls evidence exposure; method-specific internal work remains unconstrained and is charged to E2E latency.

3.2 FAILURE LOCALITY AND REPAIR SCALE

Let T_q be the units labeled unsupported before repair and n_q the answer length. Failure locality is $L_q = (\rho_q, \gamma_q, \kappa_q)$: density $\rho_q = |T_q|/n_q$; concentration $\gamma_q = \max_{r \in \mathcal{R}_q} |r|/|T_q|$ over contiguous triggered runs; and coupling κ_q, the fraction of triggers whose correction depends on an entity, temporal relation, comparison, or causal statement outside the sentence (rubric in Appendix F). All coordinates are zero when $T_q = \varnothing$.

Adjudicated initial-answer labels define T_q; verifier triggers provide a robustness check. Before outcome comparison, density and coupling were binned as low $[0, 1/3]$, medium $(1/3, 2/3]$, or high $(2/3, 1]$. The prespecified projection excludes γ_q from its interaction while retaining it to distinguish isolated from contiguous failures.

For query-level treatment effect $D_q = \mathrm{SSR}_q^{\mathrm{local}} - \mathrm{SSR}_q^{\mathrm{global}}$, we fit

$$D_q = \beta_0 + \beta_\rho d_q + \beta_\kappa c_q + \beta_{\rho\kappa} d_q c_q + \alpha_{\mathrm{dataset}(q)} + \epsilon_q,$$

where $d_q, c_q \in \{0, 1, 2\}$ encode bins, α controls dataset, and queries are equally weighted. We test $H_0 : \beta_{\rho\kappa} = 0$ by within-query local/global swaps and resample queries within dataset for intervals.

Fixed policy regions are local-favored (low/medium density, weak coupling), global-favored (high, strong), and transition (all others); Figure 2 reports their outcomes.

3.3 BUDGETED EVIDENCE PACKET COMPILATION

UNITALIGN retrieves 100 passages with Contriever-msmarco, reranks 30 spans with ms-marco-MiniLM-L12-v2, and compiles packet $\mathcal{P} \subseteq \mathcal{C}$. With inclusion x_e, $\mathrm{Rel}(e, q)$ the normalized reranker score, and $\mathrm{Tok}(e)$ the Llama-3 token count, query facets F define $\mathrm{Cov}(x, q) = |F|^{-1} \sum_{f \in F} \mathbf{1}[\exists e : x_e = 1 \wedge f \in \mathrm{facets}(e)]$; $\mathrm{Redund}(x)$ adds 0.1 per included pair with cosine similarity above 0.85.

The same Llama-3-70B-Instruct checkpoint extracts biomedical facets (population, intervention, comparator, outcome, timeframe, safety) or their cross-domain counterparts using the fixed Appendix N prompt and the shared decoding settings in Section 4.2; malformed fields are absent.

We solve the following budgeted selection problem:

$$\begin{aligned} \max_{x} \quad & \sum_{e \in \mathcal{C}} x_e \cdot \mathrm{Rel}(e, q) + \lambda \cdot \mathrm{Cov}(x, q) \\ & - \mu \cdot \mathrm{Redund}(x) \\ \text{s.t.} \quad & \sum_{e \in \mathcal{C}} x_e \cdot \mathrm{Tok}(e) \leq B \end{aligned} \tag{1}$$

with $\lambda = 2.0$ and $\mu = 0.5$, tuned on 40 held-out questions. Greedy marginal-gain selection implements Equation 1 (Algorithm 1).

3.4 GENERATION WITH SENTENCE-LEVEL PROVENANCE

The 4-bit Llama-3-70B-Instruct generator receives numbered spans `[E1]`, `[E2]`, ..., uses only packet evidence, ends every factual sentence with supporting IDs, surfaces cited conflicts, and qualifies uncertainty. Greedy decoding yields the common audited answer format.

3.5 VERIFIER-TRIGGERED LOCAL CITATION REPAIR

spaCy en_core_web_sm v3.7.4 segments sentences while preserving citations and biomedical abbreviations. DeBERTa-v3-large, trained on MNLI, SciTail, and MedNLI with source validation for checkpoint selection, scores one cited span directly or concatenated spans up to 512 tokens. Benchmark and audit pairs are excluded from training.

A separate 400-pair human set selects $\tau = 0.72$ by unsupported-class F_1, yielding 91.3% precision, 88.7% recall, 90.0% F_1, and ECE 0.038.

For $\mathcal{V}(u_i) < \tau$, a query combining s_i, missing facets, and original cited text retrieves $K_3 = 5$ spans. The generator rewrites only s_i, after which verification repeats; failure after $M = 2$ iterations adds `[UNCERTAIN]`. Algorithm 2 gives the loop.

3.6 ANSWER-WIDE COMPARISON POLICY

Both fixed-state branches receive q, the initial answer, packet, boundaries, citations, verifier scores, and T_q; both return unchanged when $T_q = \varnothing$. Otherwise, the answer-wide branch concatenates triggers into one query, retrieves K_3 spans once, and greedily regenerates every sentence and citation from the augmented packet with a 512-token cap. It is verified once and unresolved units are marked `[UNCERTAIN]`. The local branch applies Section 3.5 per trigger. Appendix N gives both prompts.

This is a comparison of complete policies, not token-matched edit spans: the shared state fixes information at assignment, while subsequent retrieval, evidence exposure, generation, and latency are measured policy effects (Table 5).

4 EXPERIMENTAL DESIGN

Three questions organize the evaluation. **Q1:** Under a shared generator, corpus, citation schema, and evidence allowance, does audit-aligned local repair improve support, answer quality, preservation, and latency? **Q2:** With every pre-repair variable fixed, when does failure locality make sentence repair preferable to answer-wide regeneration? **Q3:** Do the comparative gains survive faithful native attribution systems, complete public audits, alternative verifiers, and other generator families? BioMed-180 supplies the deepest controlled audit; public benchmarks and Qwen and Mixtral generators test transfer.

4.1 BENCHMARKS

Primary benchmark: BioMed-180. BioMed-180 contains 180 long-form questions on efficacy (45), safety (40), dosing (30), temporal updates (25), contradictions (20), and EHR-specific evidence (20), drawn from 214K PubMed Central articles, 28K ClinicalTrials.gov records, and 86K MIMIC-IV admissions. Questions require passage-localizable, multi-sentence evidence; curation removes near-duplicates, definitional prompts, under-specified patient questions, and items without retrievable support. Two clinical pharmacists write source-grounded references with oncology adjudication. The six strata resolve variation in locality, temporal alignment, and multi-span synthesis.

Secondary transfer checks. ALCE-ASQA (64), Qasper (100), and HotpotQA-Citation (120) provide complete public audits; SciQA-60, LegalQA-40, and WikiQA-50 add cross-domain checks. BioMed-public-120 is the public-only 120-question subset of BioMed-180 used for generator-family transfer. WikiQA-50 uses multi-passage Natural Questions examples (Kwiatkowski et al., 2019). Table 6 reports stratified 50-query diagnostics, and Appendix Tables 10–11 report complete-set and paired-answer results.

4.2 BASELINES

All methods use Llama-3-70B-Instruct (4-bit NF4), the same corpora, sentence-final evidence schema, and 8,192-token visible-evidence budget. Native retrieval, reflection, and regeneration remain intact and count toward E2E latency. Prompts and decoding are shared except for published policy-specific instructions; latency uses batch size 1, temperature 0, and `max_new_tokens`=512.

B1 is standard retrieve-then-generate RAG. B2 adds UNITALIGN's sentence-final citation format, and B3 scores B2 with the external NLI verifier without changing its text. B4 implements Self-RAG retrieval and self-reflection (Asai et al., 2024); B5 implements RARR's research-and-revise loop (Gao et al., 2023a); and B6 applies the complete answer-wide policy in Section 3.6. B7 constructs an entity graph from retrieved chunks and expands one hop before reranking; B8 expands three hops and reranks the enlarged candidate set, approximating standard and enhanced GraphRAG (Han et al., 2025). B4–B6 use one answer-level revision pass. B7–B8 share UNITALIGN's encoder, reranker, packet budget, and generator; only graph-based candidate expansion differs.

Baseline fidelity and scope. B1–B3 and B6 are controls, B4–B5 faithful matched-stack re-implementations, and B7–B8 controlled approximations rather than native GraphRAG replications. Table 1 therefore supports a common-stack policy comparison. Appendix Table 9 separately evaluates faithful ALCE-style, ReClaim, and verifiable fine-grained citation generators (Cao & Wang, 2024) under their native contracts.

4.3 METRICS

For query q, SSR is $\mathrm{SSR}_q = n_q^{-1} \sum_i \mathbb{1}[s_i \text{ is completely supported by the union of its cited spans}]$. CP analogously divides supported citation links by audited links within q. Multi-citation sentences are judged jointly for SSR and link by link for CP; partial support fails SSR, a factual sentence without a valid citation fails both metrics, and `[UNCERTAIN]` is scored exactly as written. Automatic and human results average per-query SSR and CP equally across represented queries. Answer quality uses ROUGE-L (Lin, 2004), BERTScore-F1 (Zhang et al., 2020), and 1–5 helpfulness. Visible tokens sum evidence tokens exposed across answer-generation and rewrite calls, excluding instructions and generated output. TTFB runs from request acceptance through prerequisite retrieval, reranking, and packet construction to the first provisional answer token; E2E ends after verification and any

repair, when the final answer is committed. Text metrics use clinician references on BioMed-180 and dataset gold answers elsewhere. Human and automatic intervals, latency, paired helpfulness, and sentence-level differences use 1,000-resample query-cluster bootstrap; paired-answer win rates use Wilson intervals over queries.

4.4 Human Evaluation Protocol

Primary citation audit. For each of the nine BioMed-180 methods, we materialize final sentence–citation pairs by query and draw a common query-stratified sample of 360 pairs, preventing long answers from dominating. Two clinical-informatics annotators (6/4 years) inspect each sentence, cited spans, and source metadata with method identity and verifier decisions hidden. Fifteen percent of pairs are independently duplicated; raw agreement is 89.6% and Cohen's $\kappa = 0.81$. A senior annotator adjudicates every disagreement before analysis. The main rows use independent method samples; point estimates and intervals weight queries equally.

Paired and holistic audits. The query-aligned UnitAlign–RARR audit pairs 360 units by query and support target, reporting wins/ties/losses, query-cluster bootstrap deltas, and a two-sided within-query permutation test. The same annotators rate all 180 answers per method for helpfulness. In the blinded whole-answer audit, they compare factual adequacy and citation trust, allow ties, and separately score factual completeness, cross-sentence coherence, and preservation of previously supported content. Complete ALCE-ASQA, Qasper, and HotpotQA-Citation audits and these paired judgments establish answer-level transfer beyond the smaller stratified diagnostics.

The protocol reviews model outputs and public or de-identified documents without participant intervention or private participant-data collection; institutional policy classifies it outside human-subjects research requiring IRB review.

Implementation. Experiments use 8× NVIDIA H100 80GB SXM5, CUDA 12.4, driver 550.54.14, PyTorch 2.2.1, vLLM 0.4.3, and Transformers 4.40.0. Retrieval uses Contriever-msmarco, ms-marco-MiniLM-L12-v2, and Faiss HNSW; DeBERTa-v3-large is fine-tuned on MNLI+SciTail+MedNLI. Calibration on 400 held-out pairs gives $\tau = 0.72$, $F_1 = 90.0\%$, and ECE $= 0.038$. Appendix C completes the protocol and compute inventory.

5 Results: A Measured Repair-Scale Crossover

Matched pre-repair states isolate the scope policy. The controlled experiment freezes the initial answer, packet, citation bindings, verifier scores, and 82 triggered units, then runs each repair policy with its scope-induced retrieval and generation workload. Local rewriting exceeds answer-wide regeneration by +6.1 [+1.4, +10.9] triggered-unit SSR points and +44.2 [+38.1, +50.5] supported-content preservation points while saving 7.01 [6.42, 7.63] s E2E. The retrieval-call and token differences are mechanisms of the policy and appear in Table 5, not hidden matching assumptions. That table also identifies both enabling operations: removing repair reduces human SSR by 7.5 points, while removing packet compilation reduces it by 6.3 points and nearly triples visible evidence.

Failure locality predicts the crossover. Table 2 stratifies the matched-state contrast by unsupported-unit density and discourse coupling computed from adjudicated initial-answer labels. These human-labeled topology counts are distinct from the 82 verifier-triggered units in Table 5. Local repair gains 8.9 SSR points for low-density, weakly coupled failures, then yields to answer-wide regeneration by 5.8 points for high-density, strongly coupled failures. In the query-weighted model of Section 3.2, the density–coupling coefficient is -4.7 [-7.9, -1.8] SSR points; a within-query policy-label permutation rejects zero interaction ($p = 0.003$). Replacing human labels with verifier triggers preserves the ordering, and empty-trigger answers follow the predefined $(0, 0, 0)$ convention. The phase regions in Figure 2 connect the support crossover to answer-level consequences: adequacy preference is +18.4 points in the 136-query local-favored region and -11.2 in the 27-query global-favored region, while local repair preserves more passed content and remains faster in all three regions. Trigger concentration records whether failures are isolated or contiguous but is not included in this prespecified two-axis interaction.

Primary human comparison. Table 1 holds the delivered sentence–citation schema, generator, corpus, and 8,192-token visible-evidence budget fixed. Among matched-stack rows B1–B6, RARR

Table 1: Main results on BioMed-180 (180 queries). All methods share the generator, corpora, and 8,192-token visible-evidence budget; method-native control work is included in E2E latency. [R] denotes matched-stack re-implementations and [A] controlled approximations. Automatic significance and bracketed human 95% intervals use query-cluster bootstrap (†: $p < 0.01$ versus the strongest matched-stack baseline). Each human row contains 360 query-stratified audited pairs, and the final panel supplies the query-aligned paired comparison.

Method	SSR (%)		CP (%)	
	Auto	Human [95% CI]	Auto	Human [95% CI]
UNITALIGN	96.1†	94.7 [92.4, 96.5]	94.8†	93.1 [90.5, 95.2]
B1: Std RAG	86.9	82.5 [78.4, 86.1]	79.8	76.4 [72.0, 80.4]
B2: +format	88.4	84.2 [80.2, 87.6]	81.2	78.1 [73.8, 82.0]
B3: +post-hoc	88.1	83.9 [79.9, 87.4]	80.6	77.8 [73.5, 81.7]
B4: Self-RAG [R]	89.2	85.8 [81.9, 89.1]	82.4	79.6 [75.3, 83.4]
B5: RARR [R]	90.6	87.1 [83.4, 90.2]	84.2	81.4 [77.2, 85.1]
B6: Ret-Revise	89.7	86.4 [82.6, 89.6]	83.2	80.6 [76.4, 84.3]
B7: GraphRAG [A]	90.4	87.2 [83.5, 90.3]	83.6	81.9 [77.8, 85.5]
B8: Enh. GraphRAG [A]	91.8	88.6 [85.1, 91.5]	85.2	83.4 [79.4, 86.8]

Method	Help. [95% CI]	ROUGE-L	BERTScore-F1	Tokens	TTFB	E2E
UNITALIGN	4.21 [4.13, 4.29]	0.412	0.876	2,460	1.24	2.86
B1: Std RAG	3.84	0.398	0.861	7,842	2.86	3.74
B2: +format	3.87	0.401	0.863	7,756	2.91	3.89
B3: +post-hoc	3.86	0.399	0.862	7,891	2.88	4.21
B4: Self-RAG [R]	3.91	0.406	0.868	6,284	3.42	6.18
B5: RARR [R]	3.94 [3.84, 4.04]	0.404	0.866	7,128	4.86	11.04
B6: Ret-Revise	3.89	0.402	0.864	7,624	4.12	9.72
B7: GraphRAG [A]	3.92	0.405	0.867	6,418	7.72	10.38
B8: Enh. GraphRAG [A]	3.96	0.408	0.869	5,892	9.14	12.86

Paired comparison, UNITALIGN vs. RARR [R]	Wins / ties / losses	Paired delta [95% CI]	Two-sided permutation
Sentence support	48/293/19	+8.1 [+5.5, +10.8] points	$p < 0.001$
Citation precision	55/289/16	+10.8 [+7.7, +13.9] points	$p < 0.001$
Helpfulness	—	+0.27 [+0.16, +0.38]	$p < 0.001$

Table 2: Local-minus-global human SSR points [query-bootstrap 95% CI]; n_q, on BioMed-180 plus HotpotQA-Citation ($n = 300$). Positive values favor local repair. Adjudicated initial-answer labels define topology before repair; for both density and coupling, low $= [0, 1/3]$, medium $= (1/3, 2/3]$, and high $= (2/3, 1]$.

Density \ coupling	Weak	Moderate	Strong
Low	+8.9 [4.8, 12.7]; 88	+6.7 [2.4, 10.6]; 31	+2.1 [-3.6, 7.4]; 13
Medium	+7.3 [3.6, 10.8]; 48	+3.8 [-0.4, 7.9]; 34	-0.9 [-5.7, 3.6]; 19
High	+3.1 [-1.8, 7.8]; 17	-1.7 [-6.0, 2.5]; 23	-5.8 [-9.9, -1.5]; 27

[R] is strongest; UNITALIGN improves human SSR from 87.1% to 94.7% and human CP from 81.4% to 93.1%, gains of 7.6 and 11.7 points, while reducing final-answer E2E latency from 11.04 s to 2.86 s. Enh. GraphRAG [A], the strongest controlled approximation, remains 6.1 SSR and 9.7 CP points behind. RARR is the principal matched-stack answer-wide policy, the GraphRAG rows test controlled graph approximations, and Table 9 reports faithful native citation systems separately.

Each main human row is an independent query-stratified 360-pair sample; the query-aligned panel provides the paired comparison. UNITALIGN records 48/293/19 support and 55/289/16 citation-precision wins/ties/losses over RARR, producing +8.1 [+5.5, +10.8] and +10.8 [+7.7, +13.9] point query-cluster bootstrap deltas. Helpfulness rises by +0.27 [+0.16, +0.38], consistent with the

ROUGE-L and BERTScore-F1 results. All policies receive matched visible answer-time evidence; method-specific internal control work remains intact and is charged to E2E latency.

Native and public comparisons. Table 9 evaluates native ALCE-style attribution, ReClaim sentence grounding, and verifiable fine-grained citation generation under their published output contracts. At the BioMed-180 operating point, UNITALIGN leads the strongest native row by 2.4 SSR and 1.7 CP points. The independent pooled-public comparison retains margins of 1.8 SSR and 1.5 CP points. Complete public audits in Table 10 establish 6.4–7.7 point SSR advantages over RARR across ALCE-ASQA, Qasper, and HotpotQA-Citation.

Sentence support transfers to whole-answer judgment. Table 11 tests holistic adequacy and citation trust on the same query sets. On BioMed-180, blinded UNITALIGN–RARR comparisons yield 109 / 44 / 27 citation-trust and 88 / 54 / 38 adequacy wins/ties/losses. Against the strongest native BioMed system, the corresponding outcomes are 95 / 50 / 35 and 82 / 55 / 43. Dataset-wise public rows sum to the 284-query pooled audit, with 146 / 61 / 77 trust and 122 / 59 / 103 adequacy wins/ties/losses. Citation trust transfers most strongly, while public adequacy disagreements concentrate in the high-coupling coherence and synthesis regime identified by the phase diagram.

The policy ordering survives stack changes. Under Qwen2.5-72B-Instruct, UNITALIGN retains a 7.3–7.6 SSR advantage over RARR. Mixtral-8x22B-Instruct-v0.1 produces 6.9/7.4 point advantages on BioMed-public-120 and Qasper. Retriever, verifier, citation schema, audit rubric, and visible-evidence budget remain fixed in both transfers. Verifier swaps keep UNITALIGN's human SSR between 93.2% and 95.2% as ECE varies from 0.031 to 0.052 (Table 4), showing that its audited support is not specific to one verifier.

6 DISCUSSION

Repair scale is a control decision. The fixed-state experiment separates two questions that aggregate citation scores conflate: whether a policy recovers support on failed units, and how much verified content it reopens in doing so. Sentence-local repair wins both comparisons for sparse, weakly coupled errors; dense, cross-sentence failures eventually reverse the support ordering. The resulting escalation rule is concrete: begin at the auditable sentence–evidence boundary, and widen intervention when failure density and discourse coupling indicate that the sentence no longer contains the information needed for correction.

Attribution becomes executable. Audit–action alignment changes provenance from an explanatory attachment into a control interface. The same unit supports a human judgment, an automated trigger, a focused retrieval query, and a preservation guarantee for surrounding text. This correspondence accounts for the joint gains in citation trust, edit containment, and latency: UNITALIGN spends new evidence and generation where the audit identifies risk while leaving unresolved support visible.

Scope and limitations. The causal comparison concerns complete policies at the evaluated operating point—English sentence-final citations, shared visible evidence, a calibrated external verifier, and the specified retrieval and generation stack—rather than edit span under artificially token-matched downstream work. The phase diagram is a fixed-state treatment-effect analysis, not an evaluated learned router, and its reported interaction uses density and discourse coupling while retaining concentration as a descriptive topology coordinate. Native systems preserve their published control flows and therefore provide comparative rather than scope-isolating evidence. Complete public audits, verifier swaps, and three generator families establish breadth within this regime; multilingual, non-sentence, and non-text attribution require an audit unit appropriate to those outputs.

7 CONCLUSION

Citation repair exhibits a measured local–global crossover. UNITALIGN makes that crossover observable by placing human inspection, verifier decisions, and selective rewriting on the delivered sentence–evidence unit under a matched visible-evidence allowance. Its local policy improves human SSR by 7.6 points and citation precision by 11.7 points over RARR on BioMed-180, preserves 44.2 additional points of passed content in the fixed-state comparison, and retains its reliability advantage across complete public audits, native systems, and three generator families. Failure locality identifies

the boundary: local-minus-global SSR moves from +8.9 points for sparse, weakly coupled failures to -5.8 points for dense, strongly coupled failures. Citation systems can therefore use provenance for more than attribution: the support structure visible to a reader can determine both where correction starts and when it must widen.

AI Use Statement

Generative AI tools assisted manuscript preparation, including copyediting, presentation revisions, naming, frontmatter wording, formatting, and consistency checks. The authors reviewed all AI-assisted revisions and remain responsible for the final content of this submission.

Ethics Statement

Data access and human review. Experiments use PubMed Central Open Access (National Library of Medicine, 2026), ClinicalTrials.gov (U.S. National Library of Medicine, n.d.), arXiv, the Caselaw Access Project, English Wikipedia, and de-identified MIMIC-IV under its PhysioNet agreement (Johnson et al., 2023); patient-linked excerpts are not redistributed. Two domain annotators reviewed public or de-identified material without participant intervention or data collection, placing the protocol outside IRB review under institutional policy. They received $25 per hour, worked at most two hours per session, and took 15-minute breaks.

Reproducibility Statement

The protocol, models, hyperparameters, compute environment, and evaluation estimators are specified in Sections 3–4. Appendix B documents benchmark construction and evidence units; Appendices F–N provide the audit rubric, label definitions, and prompts; and Tables 9–12 report native-comparator, answer-level, and cross-stack checks.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511.

Shuyang Cao and Lu Wang. Verifiable generation with subsentence-level fine-grained citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15584–15596, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.920. URL https://aclanthology.org/2024.findings-acl.920/.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, June 2021. doi:10.18653/v1/2021.naacl-main.365.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

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

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao,

Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. Retrieval-augmented generation with graphs (GraphRAG). *arXiv preprint arXiv:2501.00309*, 2025. URL https://arxiv.org/abs/2501.00309.

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*, pp. 874–880, Online, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.495.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi:10.1038/s41597-022-01899-x.

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 Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.550.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. In Lillian Lee, Mark Johnson, Kristina Toutanova, and Brian Roark (eds.), *Transactions of the Association for Computational Linguistics*, volume 7, pp. 453–466, March 2019. doi:10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

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, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Harsh Maheshwari, Srikanth Tenneti, and Alwarappan Nakkiran. CiteFix: Enhancing RAG accuracy through post-processing citation correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 310–317, Vienna, Austria, July 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.acl-industry.23. URL https://aclanthology.org/2025.acl-industry.23/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.557.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali

(eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.741.

National Library of Medicine. PMC open access subset, 2026. URL https://pmc.ncbi.nlm.nih.gov/tools/openftlist/. Last modified: 2026-08-24.

Yixing Peng, Licheng Zhang, Shancheng Fang, Yi Liu, Peijian Gu, and Quan Wang. FineRef: Fine-grained error reflection and correction for long-form generation with citations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32692–32700, 2026. doi:10.1609/aaai.v40i39.40547. URL https://doi.org/10.1609/aaai.v40i39.40547.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. VeriCite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 47–54, Xi'an, China, 2025. Association for Computing Machinery. doi:10.1145/3767695.3769505. URL https://doi.org/10.1145/3767695.3769505.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023. doi:10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

U.S. National Library of Medicine. ClinicalTrials.gov, n.d. URL https://clinicaltrials.gov/. Accessed: 2024-08.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.609.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi:10.18653/v1/2022.findings-emnlp.347.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76, Seattle, United States, July 2022b. Association for Computational Linguistics. doi:10.18653/v1/2022.findings-naacl.6.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 969–988, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.findings-naacl.55. URL https://aclanthology.org/2025.findings-naacl.55/.

Zhichao Yan, Shizhao Li, Jiapu Wang, Haoran Luo, Qingang Zhang, Jiaoyan Chen, Ru Li, and Jeff Z. Pan. CAGE: Cognitive attribution graphs for faithful inline citation generation in long-form question answering. *arXiv preprint arXiv:2607.24236*, 2026. URL https://arxiv.org/abs/2607.24236.

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, October 2018. doi:10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Operational Algorithms

Algorithm 1 Budgeted Evidence Packet Compilation

Require: Query q, candidates $\mathcal{C}$, budget B, weights λ, μ
Ensure: Evidence packet $\mathcal{P}$
$\mathcal{P} \leftarrow \emptyset$; tokens $\leftarrow 0$
Compute $\mathrm{Rel}(e, q)$ for all $e \in \mathcal{C}$
Extract facets F from q via few-shot prompting
while $\mathcal{C} \neq \emptyset$ **do**
  for $e \in \mathcal{C}$ **do**
    $\Delta_e \leftarrow \mathrm{Rel}(e, q) + \lambda \cdot \Delta\mathrm{Cov}(e) - \mu \cdot \Delta\mathrm{Redund}(e)$
  end for
  $e^* \leftarrow \arg\max_e \Delta_e$
  if tokens $+ \mathrm{Tok}(e^*) > B \vee \Delta_{e^*} \leq 0$ **then**
    break
  end if
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{e^*\}$; tokens $\leftarrow$ tokens $+ \mathrm{Tok}(e^*)$
  $\mathcal{C} \leftarrow \mathcal{C} \setminus \{e^*\}$
end while
return $\mathcal{P}$

Algorithm 2 Verifier-Triggered Local Citation Repair

Require: Answer A, packet $\mathcal{P}$, verifier $\mathcal{V}$, threshold τ, max iterations M
Ensure: Repaired answer A'
$A' \leftarrow A$
$\{s_1, \ldots, s_n\} \leftarrow \mathrm{Segment}(A')$
for $i = 1$ to n **do**
  iter $\leftarrow 0$
  while $\mathcal{V}(s_i, \mathrm{cited}(s_i)) < \tau \wedge \mathrm{iter} < M$ **do**
    $q_{\mathrm{ret}} \leftarrow \mathrm{BuildQuery}(s_i, \mathrm{missing_facets}, \mathrm{cited}(s_i))$
    $\mathcal{E}_{\mathrm{new}} \leftarrow \mathrm{Retrieve}(q_{\mathrm{ret}}, K_3)$
    $s_i' \leftarrow \mathrm{Rewrite}(s_i, \mathcal{P} \cup \mathcal{E}_{\mathrm{new}})$
    $s_i \leftarrow s_i'$; iter $\leftarrow$ iter $+ 1$
  end while
  if $\mathcal{V}(s_i, \cdot) < \tau$ **then**
    $s_i \leftarrow s_i \oplus$ [UNCERTAIN]
  end if
  Update A' with s_i
end for
return A'

B Benchmark Details

ALCE-ASQA. 64 ambiguous questions from ALCE (Gao et al., 2023b). Evidence: English Wikipedia (Aug 2024), 6.2M passages. Gold citations from original annotations.

Qasper. 100 questions from Qasper (Dasigi et al., 2021) requiring synthesis across paper sections. Evidence: full-text PDFs converted to 300-word passages. 47 papers, avg 42 passages/paper.

HotpotQA-Citation. 120 multi-hop questions from HotpotQA dev set (Yang et al., 2018). Evidence: 10 distractor paragraphs + 2 gold paragraphs per question. Citation binding evaluated against supporting facts.

Cross-domain sets. SciQA-60: 60 NLP/ML questions on 47,832 arXiv papers (cs.CL, cs.LG, stat.ML). LegalQA-40: 40 questions on 12,418 U.S. court opinions (2015–2023). WikiQA-50: 50 Natural Questions requiring 3+ passages. These cross-domain slices map the shared facet scaffold from Section 3.3 to domain-appropriate entities, actions, outcomes, temporal scope, and exceptions, testing one selection policy across scientific, legal, and encyclopedic evidence.

C Protocol and Compute

Protocol surface. The appendix specifies prompts, retriever and verifier settings, calibration, audit rules, and estimator semantics. Together with the model versions and hardware in Section 4.4, these details define the decisions needed to interpret the reported comparisons.

Compute accounting. Per-query wall-clock measurements separate retrieval and reranking, packet compilation, generation, verification, and targeted rewriting. The main inference sweep evaluates nine Table 1 systems on 180 BioMed-180 queries; transfer experiments cover the public and cross-domain sets in Tables 6 and 8. One-time training consists of DeBERTa-v3-large fine-tuning on MNLI, SciTail, and MedNLI, followed by threshold calibration on 400 held-out pairs.

D BioMed-180 Construction and Audit Scope

BioMed-180 is a high-resolution stress test for citation-sensitive long-form answers. Its concentration on temporal updates, contradictions, dosing details, and EHR-linked synthesis creates controlled variation in unsupported-unit density and discourse coupling. Full public ALCE-ASQA, Qasper, and HotpotQA-Citation evaluations complement this diagnostic operating point with standard evidence structures and faithful native comparators.

Annotators judge each delivered sentence with the evidence spans it explicitly cites, matching the object a user inspects. Support labels use only visible bindings; helpfulness and paired adequacy independently measure holistic utility. This separation lets the experiments identify whether a change improves evidential support, answer quality, or both.

E Baseline Instantiation Notes

Comparison scope. B1–B3 and Retrieve-Revise are controlled baselines; Self-RAG and RARR are matched-stack re-implementations; GraphRAG variants are controlled approximations. This partition lets Table 1 isolate policy families with a common generator, corpus, evidence budget, and output schema. Table 9 then provides the complementary faithful-system comparison for ALCE, ReClaim, and verifiable fine-grained citations under native control and output contracts.

Row-level control flow. Every matched-stack row starts from Contriever top-100 retrieval and MiniLM top-30 reranking. B1 generates once from ranked spans; B2 uses the same path with the sentence-final citation instruction; and B3 applies the calibrated NLI score to B2 without retrieval or revision. B4 exposes the shared top-30 spans to the Self-RAG reflection controller and permits one answer-level revision. B5 uses RARR's claim research and revision sequence, retrieving up to five additional spans per research request before one revised answer. B6 concatenates the triggered sentences into one query, retrieves five spans once, and executes the answer-wide prompt below. B7 and B8 expand the initial entity graph by one and three hops, respectively, rerank the resulting candidates, and generate once. Local UnitAlign instead issues at most one five-span retrieval per failed unit per iteration and stops after support passes or $M = 2$. All generation calls use greedy decoding and a 512-token answer cap unless a native system's published controller requires a different termination rule.

F Annotator Instructions

The annotation interface applies the following decision rules to every audited sentence–citation pair:

1. Judge one delivered sentence together with the evidence spans it explicitly cites; do not use outside knowledge or uncited context to rescue a sentence.
2. Assign "Supported" only if the cited span(s) support the full sentence as written, including quantities, qualifiers, and temporal scope.
3. Use "Partial," "Mis-cited," "Multi-span," "Temporal," and "Other" according to the label definitions below; uncertain cases are flagged for adjudication rather than guessed.
4. Rate helpfulness separately from support: a useful answer can still be unsupported, and a fully supported answer can still be unhelpful.
5. Read only the sentence, cited spans, and source metadata shown in the interface; system identity remains hidden throughout annotation.

A sentence ending with `[UNCERTAIN]` is still judged under the same label definitions as written; the flag does not invoke a separate audit rubric. For the failure-locality analysis, annotators separately mark discourse coupling when correcting the unit requires an entity, temporal relation, comparison, causal link, or evidential qualification established outside that sentence. This diagnostic tag can co-occur with any support label and does not change SSR or CP.

Whole-answer comparison rubric. For each paired query, annotators choose system A, system B, or tie separately for factual adequacy and citation trust. Adequacy rewards complete, correct resolution of the question; citation trust rewards claims whose attached evidence can be followed and supports the text as written. Annotators also score factual completeness and cross-sentence coherence from 1 (substantially deficient) to 5 (complete or coherent throughout), and mark whether content supported in the initial answer remains supported. Pair order is randomized, system identity is hidden, and a senior annotator adjudicates disagreements before win/tie/loss aggregation.

G Human Audit Label Definitions

- Supported: Cited span(s) entail the sentence.
- Partial: Main claim is supported but a qualifier, quantity, or scope condition is not; this label fails SSR.
- Mis-cited: Topically related but not supporting.
- Multi-span: Support requires combining spans. The unit passes SSR only when their visible union entails the complete sentence.
- Temporal: Different time period than stated.
- Other: Unsupported for other reasons.

H Sensitivity Analysis

Table 3 shows sensitivity to key hyperparameters.

The budget sweep shows diminishing returns beyond the default setting. Moving from 8,192 to 16,384 tokens improves human SSR only slightly while increasing TTFB noticeably, so the default budget already captures most of the attainable support gain. Reducing the budget to 4,096 is more damaging than changing the packet weights within a modest range, which indicates that evidence headroom is a real constraint rather than a cosmetic tuning knob. The coverage reward also matters more than the redundancy penalty in this regime because under-covered questions are harder to repair than mildly redundant packets. Finally, targeted retrieval beyond five extra spans adds little additional support relative to its latency cost, which is why we keep $K_3 = 5$ as the default.

I Verifier Robustness

Table 4 shows results with alternative verifiers.

Table 3: Sensitivity analysis on BioMed-180. Each block varies one hyperparameter while holding the remaining defaults fixed; SSR and CP are human percentages and TTFB is the median in seconds.

Parameter	Value	SSR	CP	TTFB
Budget B	4,096	93.2	91.4	0.94
	8,192 (default)	94.7	93.1	1.24
	16,384	95.1	93.8	1.86
λ (coverage)	1.0	92.4	89.8	1.21
	2.0 (default)	94.7	93.1	1.24
	3.0	94.2	92.6	1.28
μ (redundancy)	0.25	93.8	91.2	1.22
	0.5 (default)	94.7	93.1	1.24
	1.0	93.4	92.8	1.31
K_3 (targeted)	3	93.6	91.8	1.19
	5 (default)	94.7	93.1	1.24
	10	94.9	93.4	1.38

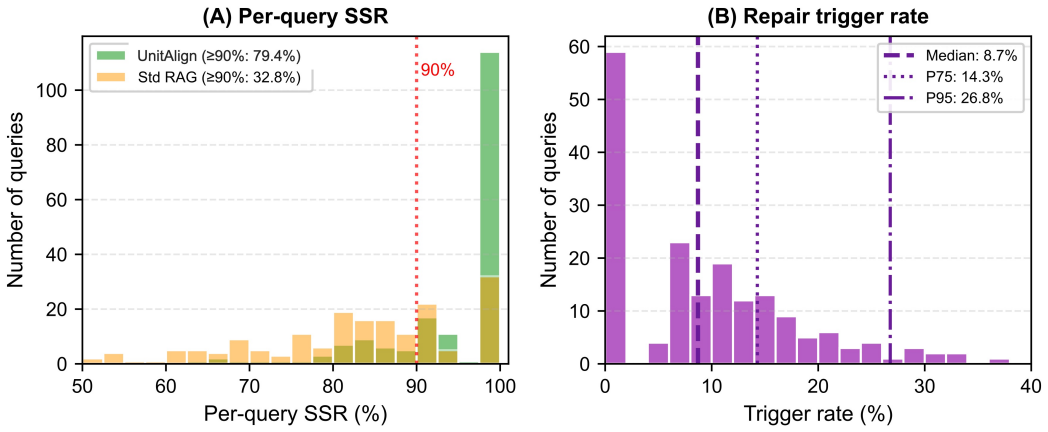
Table 4: Verifier robustness on BioMed-180. Each row changes only the verifier; SSR and CP are human percentages, concordance measures agreement with adjudicated support labels, and ECE measures calibration.

Verifier	SSR	CP	Concordance	ECE
DeBERTa-v3-large (default)	94.7	93.1	93.1%	0.038
RoBERTa-large-MNLI	93.2	91.6	89.4%	0.052
Llama-3-8B-judge	93.8	92.4	90.8%	0.046
Ensemble (3 verifiers)	95.2	93.8	94.6%	0.031

Human SSR remains above 93% under all four verifiers. The ensemble reaches the strongest scores, while its modest gain over DeBERTa-v3-large shows that the default result is representative rather than verifier-selected. RoBERTa's lower concordance and higher ECE translate into more missed and unnecessary rewrites, confirming that calibration affects control quality as well as offline ranking.

J Additional Transfer and Efficiency Results

Aligned audit boundary. The scientific object is one delivered sentence together with the evidence spans it explicitly cites. Annotators judge whether those spans support the complete sentence as written; the verifier receives the same unit; and unsupported units define the minimum retrieval-and-rewrite target. Trigger-rate, ablation, and latency diagnostics consequently share one observation and control boundary, while Table 11 supplies the complementary whole-answer view of adequacy and citation trust.



Policy region	n_q	Adequacy preference / preservation / E2E deltas
Local-favored	136	+18.4 / +43.8 / -6.90
Transition	137	+5.7 / +29.4 / -5.42
Global-favored	27	-11.2 / +12.7 / -2.10

Figure 2: Distribution and failure-locality evidence. Panel (A) shows per-query SSR and marks 90%; panel (B) shows per-query trigger rates with median, P75, and P95. The 8.3% value in the text is a sentence-weighted mean, whereas the 8.7% marker is the median per-query rate. The matrix reports local-minus-global adequacy-preference points, supported-content preservation points, and E2E seconds by policy region; all are query-level deltas. Table 2 reports the complete density–coupling SSR matrix and query-cluster intervals.

The distribution shift spans the BioMed-180 query set rather than concentrating in a few easy items. The policy regions summarize the boundary established in Table 2: preservation remains positive even in the global-favored corner, but holistic adequacy reverses as repair requires more cross-sentence coordination.

Table 5: Component ablations and fixed-state repair policies on BioMed-180. In the policy panels, all rows receive the same initial answers, packet, sentence boundaries, citations, verifier scores, and trigger set; "Trig." evaluates the 82 triggered units, "Pres." is supported-content preservation outside those units, and "Edit" is normalized edit distance outside them. Retrieval calls and visible tokens are policy outcomes.

Configuration	SSR	CP	Tokens	TTFB
Full UNITALIGN	94.7	93.1	2,460	1.24
w/o repair	87.2	82.8	2,520	1.18
w/o packet	88.4	87.9	7,140	1.93
w/o multi-span	93.2	91.4	2,480	1.21
w/o coverage	91.8	86.3	2,680	1.29

Policy	Trig. SSR / CP	Full SSR / CP	Pres.	Edit
No repair	31.7 / 28.0	87.2 / 82.8	100.0	0.0
Sentence-local	90.2 / 87.8	94.7 / 93.1	98.9	0.8
Whole-answer	84.1 / 81.7	91.5 / 88.6	54.7	31.6

Policy	Help. / coherence	Retrieval calls / tokens	TTFB / E2E (s)
No repair	4.03 / 4.09	0.00 / 2,520	1.18 / 2.49
Sentence-local	4.21 / 4.18	0.46 / 2,460	1.24 / 2.86
Whole-answer	4.13 / 4.06	1.00 / 7,130	4.31 / 9.87

Table 5 identifies two complementary effects. Removing local repair produces the largest reliability drop; removing packet compilation primarily sacrifices evidence efficiency. Under the stronger matched-state control (180 / 987 / 82: queries / initial units / triggered units), sentence-local repair exceeds whole-answer regeneration by +6.1 [+1.4, +10.9] triggered-unit SSR points and +44.2 [+38.1, +50.5] preservation points, with query-cluster 95% intervals. It also reduces E2E latency by 7.01 [6.42, 7.63] s. The estimand is the total effect of selecting a repair scope, including the retrieval and generation work that scope induces after the shared pre-repair state.

Table 6: Small transfer audits on stratified 50-query development subsets. SSR/CP are query-weighted human percentages with query-cluster 95% intervals; Help. is mean answer helpfulness.

Benchmark	Method	SSR (%)	CP (%)	Help.
ALCE-ASQA	UNITALIGN	91.8 [87.2, 95.1]	89.4 [84.6, 93.2]	4.08
	Self-RAG	83.6 [78.1, 88.2]	80.2 [74.8, 84.9]	3.86
	RARR	84.9 [79.6, 89.3]	81.6 [76.4, 86.1]	3.89
Qasper	UNITALIGN	93.2 [88.9, 96.4]	91.6 [87.1, 95.0]	4.14
	Self-RAG	82.4 [76.8, 87.2]	78.8 [73.1, 83.6]	3.81
	RARR	85.1 [79.8, 89.5]	82.4 [76.9, 87.0]	3.91
HotpotQA	UNITALIGN	92.6 [88.4, 95.8]	90.8 [86.2, 94.3]	4.11
	Self-RAG	81.8 [76.1, 86.7]	78.2 [72.4, 83.2]	3.78
	RARR	84.2 [78.8, 88.7]	81.4 [75.9, 86.1]	3.87

The ordering is consistent across ambiguous, scientific-document, and multi-hop questions. Tables 10 and 11 increase both the public sample size and evaluation granularity.

Latency mechanism. Figure 3 separates first-token responsiveness from full-completion cost. Generation dominates; packet compilation, verification, and sparse rewriting account for the remaining E2E latency after streaming begins.

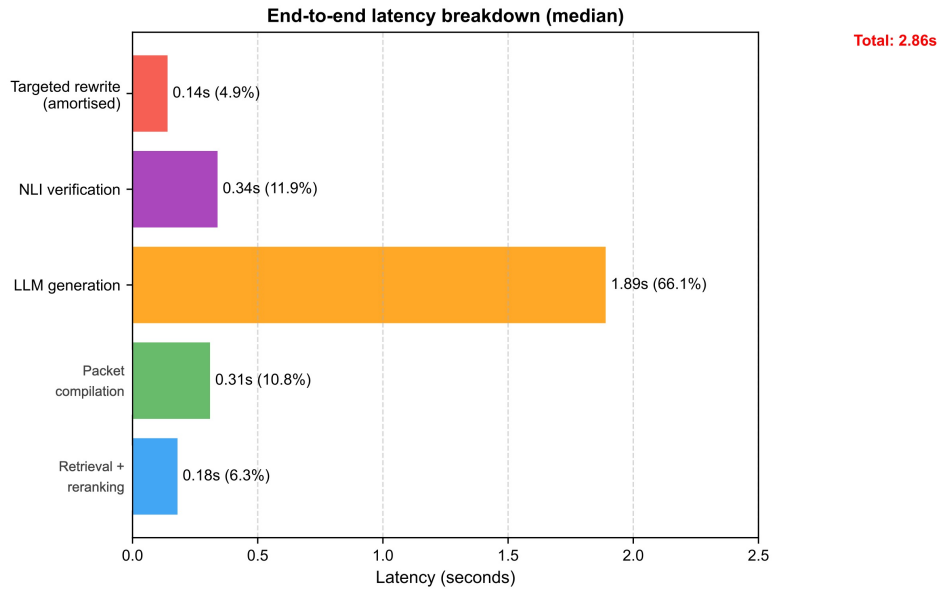


Figure 3: Latency breakdown (median over 180 queries). Retrieval and reranking, packet compilation, generation, verification, and amortized targeted rewrite sum to the 2.86 s E2E median in Table 1. TTFB is smaller because verification and rewrite continue after streaming begins.

K RETRIEVAL DIAGNOSTICS

Table 7 shows retrieval recall@k across domains. Recall@k uses benchmark-provided passage-level gold evidence for BioMed-180, ALCE-ASQA, Qasper, HotpotQA-Citation, and LegalQA-40. Human SSR and CP provide the evidence diagnostic for SciQA-60 and WikiQA-50.

Table 7: Retrieval recall@k using benchmark-provided passage-level gold evidence. Trigger rate is the sentence-weighted fraction of units selected for repair.

Domain	R@10	R@30	R@100	Trigger rate
BioMed-180	78.4	89.2	94.6	8.3%
ALCE-ASQA	82.1	91.4	96.2	7.1%
Qasper	74.6	86.8	93.1	9.8%
HotpotQA	81.4	90.6	95.4	7.6%
LegalQA	68.2	81.4	89.6	12.4%

LegalQA shows lower retrieval recall (68.2% R@10), explaining its lower SSR gains. Targeted retrieval compensates partially (trigger rate 12.4% vs 8.3% for BioMed). The broader pattern is that higher trigger rates track lower first-pass recall, which is exactly the failure mode a repair loop should expose. BioMed-180, ALCE-ASQA, and HotpotQA all have strong R@100, so local repair mostly needs to recover from sentence-level support mistakes rather than complete retrieval misses. LegalQA is different because the evidence gap starts earlier in the pipeline. Table 7 therefore helps separate retrieval failure from citation-binding failure.

L EXTENDED CROSS-DOMAIN RESULTS

Table 8: Extended cross-domain consistency checks. SSR/CP are query-weighted human percentages with query-cluster 95% intervals; Help. is mean answer helpfulness.

Domain	Method	SSR (%)	CP (%)	Help.
SciQA-60	UNITALIGN	92.4 [88.1, 95.8]	90.8 [86.2, 94.3]	4.12
	Self-RAG	82.8 [77.2, 87.5]	79.4 [73.6, 84.4]	3.84
	RARR	84.6 [79.2, 89.1]	81.2 [75.6, 85.9]	3.89
	Std RAG	79.6 [73.8, 84.6]	74.2 [68.1, 79.6]	3.76
LegalQA-40	UNITALIGN	89.2 [83.6, 93.4]	87.4 [81.4, 92.1]	3.98
	Self-RAG	80.4 [73.8, 85.8]	77.2 [70.6, 82.8]	3.72
	RARR	82.1 [75.8, 87.2]	79.6 [73.2, 85.0]	3.78
	Std RAG	78.4 [71.2, 84.3]	73.6 [66.4, 79.8]	3.64
WikiQA-50	UNITALIGN	93.8 [89.4, 96.9]	92.1 [87.5, 95.4]	4.18
	Self-RAG	84.2 [78.6, 88.9]	81.4 [75.6, 86.2]	3.92
	RARR	86.4 [81.0, 90.8]	83.8 [78.2, 88.4]	3.96
	Std RAG	81.2 [74.8, 86.4]	76.8 [70.4, 82.2]	3.82

The ordering persists across scientific, legal, and general-reference domains. LegalQA is hardest and has the weakest retrieval recall in Table 7, connecting reduced citation support to an upstream evidence-access bottleneck.

M NATIVE, PUBLIC, AND CROSS-STACK VALIDATION

Unless native control requires stochastic search, inference uses temperature 0 and seed 17. Intrinsically stochastic native systems use seeds 17/18/19 and the median query-level metric before auditing. Human SSR/CP, helpfulness, and latency use 1,000 query-cluster bootstrap resamples; paired full-answer win rates use Wilson intervals over shared queries.

M.1 NATIVE CLOSEST-BASELINE CLOSURE

Table 9: Native closest-baseline closure under faithful output and control contracts. The public panel pools the 50-query ALCE-ASQA, Qasper, and HotpotQA-Citation slices. The BioMed-180 panel uses the main corpus, 8,192-token visible-evidence allowance, Llama-3-70B-Instruct when a shared generator is permitted, and the same 360-pair query-stratified audit design for every method. Human SSR/CP are query-weighted means with query-cluster 95% intervals. Count panels report pooled sentence pairs audited / supported / unsupported; CP uses a separate citation-link denominator, so count-derived rates need not equal weighted estimates.

Pooled public method	Human SSR [95% CI]	Human CP [95% CI]	Help.
UNITALIGN	92.5 (90.3–94.3)	90.6 (88.1–92.7)	4.11
RARR [R]	84.7 (81.9–87.2)	81.8 (78.8–84.4)	3.89
Native ALCE-style	88.1 (85.1–90.6)	85.6 (82.4–88.2)	3.97
Native ReClaim	89.9 (87.0–92.2)	87.8 (84.8–90.4)	4.00
Native verifiable citation	90.7 (88.0–92.9)	89.1 (86.4–91.5)	4.04

Pooled public method	Tokens / E2E	TTFB	n_q; audited / supported / unsupported
UNITALIGN	2,740 / 3.18	1.39	150; 694 / 642 / 52
RARR [R]	6,410 / 8.94	4.63	150; 694 / 588 / 106
Native ALCE-style	5,980 / 4.41	3.06	150; 676 / 596 / 80
Native ReClaim	5,120 / 5.02	3.24	150; 681 / 612 / 69
Native verifiable citation	4,380 / 4.86	2.71	150; 687 / 623 / 64

BioMed-180 method	Human SSR [95% CI]	Human CP [95% CI]	Help.
UNITALIGN	94.7 [92.4, 96.5]	93.1 [90.5, 95.2]	4.21
RARR [R]	87.1 [83.4, 90.2]	81.4 [77.2, 85.1]	3.94
Native ALCE-style	89.2 [85.6, 92.1]	86.7 [82.8, 89.8]	4.03 [3.94, 4.12]
Native ReClaim	91.6 [88.4, 94.0]	90.2 [86.7, 92.9]	4.13 [4.05, 4.21]
Native verifiable citation	92.3 [89.2, 94.6]	91.4 [88.1, 93.8]	4.16 [4.08, 4.24]

BioMed-180 method	Tokens	TTFB	E2E	n_q; audited / supported / unsupported
UNITALIGN	2,460	1.24	2.86	180; 360 / 341 / 19
RARR [R]	7,128	4.86	11.04	180; 360 / 314 / 46
Native ALCE-style	6,010	3.14	4.58	180; 360 / 321 / 39
Native ReClaim	5,290	3.31	5.42	180; 360 / 330 / 30
Native verifiable citation	4,510	2.78	4.96	180; 360 / 332 / 28

The pooled public comparison preserves the previously observed native ordering. On BioMed-180, verifiable fine-grained citation generation is the strongest native competitor, and UNITALIGN retains 2.4 SSR and 1.7 CP point advantages with lower TTFB and E2E latency. The consistent public and same-setting results establish that audit-unit repair improves over both answer-wide revision and generation-time citation grounding.

M.2 Larger Public Human Validation and Paired Full-Answer Audit

Table 10: Complete public human audits with the main expert sentence–citation rubric. Human SSR/CP are query-weighted means with query-cluster 95% intervals. Counts pool sentence pairs audited / supported / unsupported, while CP uses a separate citation-link denominator.

Dataset	Method	Human SSR (%)	Human CP (%)
ALCE-ASQA (64)	UnitAlign	90.6 (86.6–93.8)	88.8 (84.3–92.3)
	RARR [R]	84.2 (79.4–88.2)	81.0 (76.0–85.2)
	Self-RAG [R]	82.7 (77.7–86.9)	79.1 (74.0–83.5)
Qasper (100)	UnitAlign	92.1 (89.0–94.4)	90.5 (87.1–93.0)
	RARR [R]	84.8 (81.1–88.0)	82.0 (78.1–85.4)
	Self-RAG [R]	82.4 (78.5–85.8)	78.7 (74.6–82.4)
HotpotQA-Citation (120)	UnitAlign	91.4 (88.6–93.7)	89.9 (86.8–92.4)
	RARR [R]	83.7 (80.4–86.8)	80.8 (77.2–84.0)
	Self-RAG [R]	81.5 (78.0–84.7)	78.0 (74.3–81.4)

Dataset	Method	Help.	Audited / supported / unsupported
ALCE-ASQA (64)	UnitAlign	4.04	278 / 252 / 26
	RARR [R]	3.90	278 / 234 / 44
	Self-RAG [R]	3.84	278 / 230 / 48
Qasper (100)	UnitAlign	4.09	421 / 388 / 33
	RARR [R]	3.93	421 / 357 / 64
	Self-RAG [R]	3.82	421 / 347 / 74
HotpotQA-Citation (120)	UnitAlign	4.05	503 / 460 / 43
	RARR [R]	3.88	503 / 421 / 82
	Self-RAG [R]	3.79	503 / 410 / 93

Across all three complete public sets, UnitAlign improves SSR by 6.4–7.7 points over RARR and also increases citation precision and helpfulness. The shared direction across ambiguous, full-document, and multi-hop questions supports transfer beyond BioMed-180.

Table 11: Blinded paired full-answer audit. Adequacy and citation-trust W/T/L labels are mutually exclusive primary outcomes on each query. Preference direction is determined by wins versus losses; win-rate columns report wins over all paired queries, with Wilson 95% intervals. Dataset-wise public rows sum to the pooled 284-query row. Diagnostic failure tags can co-occur. The lower panel reports query-paired mean score deltas and preservation of initially supported content on BioMed-180.

Comparison set	n_q	Adequacy W/T/L	Adeq. win rate	Citation trust W/T/L	Trust win rate
BioMed-180: UNITALIGN vs. RARR	180	88 / 54 / 38	48.9 (41.7–56.1)	109 / 44 / 27	60.6 (53.3–67.4)
BioMed-180: UNITALIGN vs. native verifiable citation	180	82 / 55 / 43	45.6 (38.4–52.8)	95 / 50 / 35	52.8 (45.5–59.9)
ALCE-ASQA	64	31 / 13 / 20	48.4 (36.6–60.4)	35 / 15 / 14	54.7 (42.6–66.3)
Qasper	100	40 / 21 / 39	40.0 (30.9–49.8)	51 / 20 / 29	51.0 (41.3–60.6)
HotpotQA-Citation	120	51 / 25 / 44	42.5 (34.0–51.4)	60 / 26 / 34	50.0 (41.2–58.8)
Public pooled	284	122 / 59 / 103	43.0 (37.3–48.8)	146 / 61 / 77	51.4 (45.6–57.2)

BioMed-180 comparison	Factual completeness Δ	Cross-sentence coherence Δ	Supported-content preservation
UNITALIGN vs. RARR	+0.21 [+0.11, +0.31]	+0.08 [-0.01, +0.17]	96.8 [94.1, 98.3]
UNITALIGN vs. native verifiable citation	+0.12 [+0.03, +0.21]	+0.04 [-0.05, +0.13]	93.6 [89.8, 96.0]

On the primary benchmark, citation-trust wins outnumber losses four to one against RARR, while adequacy wins exceed losses by 50 queries and supported-content preservation reaches 96.8 [94.1, 98.3]. The public rows show a smaller but consistent trust advantage across all three datasets. Among the 162 pooled public adequacy non-wins, recurring diagnostic tags are cross-sentence coherence 23 / 162 (14.2 (9.6–20.4)), under-specification 19 / 162 (11.7 (7.7–17.6)), and missed cross-paragraph synthesis 14 / 162 (8.6 (5.2–13.9)). These tags connect holistic outcomes to the density–coupling policy boundary in Figure 2.

M.3 GENERATOR-ARCHITECTURE ROBUSTNESS

We replace Llama-3-70B-Instruct with Qwen2.5-72B-Instruct and the mixture-of-experts Mixtral-8x22B-Instruct-v0.1 while holding the retriever, verifier, visible-evidence budget, audit rubric, and bracketed citation schema fixed. Each generator receives only the short adapter needed to express the shared citation format.

Table 12: Generator-architecture robustness on BioMed-public-120 and Qasper. Human SSR/CP are query-weighted means with query-cluster 95% intervals; the final panel reports pooled sentence-pair counts. Tokens, TTFB, and E2E are medians.

Generator	Dataset	Method	Human SSR (%)	Human CP (%)	Help.
Qwen2.5-72B	BioMed-public	UNITALIGN	91.9 (88.2–94.7)	90.2 (86.2–93.2)	4.02
		RARR [R]	84.6 (79.7–88.5)	81.7 (76.7–85.9)	3.88
		Std RAG	80.4 (75.1–84.9)	75.8 (70.2–80.8)	3.74
	Qasper	UNITALIGN	90.8 (87.7–93.3)	89.0 (85.6–91.8)	4.01
		RARR [R]	83.2 (79.3–86.6)	80.3 (76.0–84.1)	3.85
		Std RAG	78.9 (74.7–82.6)	73.8 (69.2–78.0)	3.68
Mixtral-8x22B	BioMed-public	UNITALIGN	90.8 (87.0–93.7)	89.3 (85.2–92.5)	3.98
		RARR [R]	83.9 (79.0–87.9)	80.8 (75.6–85.2)	3.84
		Std RAG	79.6 (74.3–84.0)	74.9 (69.3–79.8)	3.70
	Qasper	UNITALIGN	89.9 (86.5–92.6)	88.2 (84.5–91.1)	4.00
		RARR [R]	82.5 (78.4–86.0)	79.7 (75.3–83.6)	3.83
		Std RAG	78.2 (73.8–82.1)	72.9 (68.1–77.3)	3.65

Generator	Dataset	Method	Tokens / TTFB / E2E / audited / supported / unsupported
Qwen2.5-72B	BioMed-public	UNITALIGN	2,730 / 1.58 / 3.44 / 262 / 241 / 21
		RARR [R]	6,980 / 5.07 / 10.96 / 262 / 222 / 40
		Std RAG	7,540 / 3.22 / 4.27 / 262 / 211 / 51
	Qasper	UNITALIGN	2,860 / 1.66 / 3.57 / 417 / 379 / 38
		RARR [R]	6,240 / 4.88 / 9.73 / 417 / 347 / 70
		Std RAG	7,180 / 3.08 / 4.11 / 417 / 329 / 88
Mixtral-8x22B	BioMed-public	UNITALIGN	2,810 / 1.73 / 3.71 / 265 / 241 / 24
		RARR [R]	6,740 / 5.18 / 10.42 / 265 / 222 / 43
		Std RAG	7,390 / 3.34 / 4.48 / 265 / 211 / 54
	Qasper	UNITALIGN	2,920 / 1.79 / 3.84 / 414 / 372 / 42
		RARR [R]	6,310 / 5.03 / 9.89 / 414 / 342 / 72
		Std RAG	7,240 / 3.19 / 4.26 / 414 / 324 / 90

Absolute support varies by generator, but both dense Qwen and mixture-of-experts Mixtral preserve the local-over-global ordering. UNITALIGN exceeds RARR by 7.3/7.6 SSR points under Qwen and 6.9/7.4 under Mixtral on BioMed-public/Qasper, while retaining substantially lower TTFB and E2E latency.

N PROMPT TEMPLATES

The templates below define the common citation schema and the two intervention scopes used in the fixed-state comparison.

> *Facet extraction.* Extract a compact query record with these fields: population, intervention, comparator, outcome, time, risks. Use only information stated in the query. Return an

empty value for an absent field; do not infer one. For non-biomedical questions, map the fields to subject/entity, action or alternative, result/holding, temporal scope, and risks or exceptions. Return only the field names and values.

Evidence packet generation. You are given evidence spans [E1], [E2], ...
Answer using ONLY the provided evidence.
End EVERY sentence with citation(s) like [E1] or [E1][E3].
If evidence conflicts, state both views; if uncertain, say so.

Whole-answer repair. Question: [query].
Initial answer: [answer].
Sentences flagged as not fully supported: [triggered_sentences].
Evidence: [packet] [new_spans].
Regenerate the COMPLETE answer using ONLY the supplied evidence. Resolve the flagged support problems and preserve correct information when evidence permits. Every factual sentence must end with citation(s) such as [E1] or [E1][E3]. If the evidence does not support a claim, omit it or mark it [UNCERTAIN].

Targeted rewrite. The following sentence was flagged as not fully supported: [sentence].
Evidence: [spans] [new_spans].
Rewrite ONLY this sentence to be fully supported. Keep its meaning and end with citations.

O QUALITATIVE EXAMPLES

Example 1: Mis-citation. For "Metformin should be avoided in eGFR $<$30 [E4]," [E4] discusses dose adjustment, whereas focused retrieval adds the contraindication threshold in [E7]. UNITALIGN writes "Metformin is contraindicated in eGFR $<$30 mL/min/1.73m² [E4][E7]." Entailment rises from 0.42 to 0.91; the adjacent monitoring and dose-adjustment sentences remain byte-identical.

Example 2: Temporal mismatch. For "Trastuzumab plus chemotherapy is standard [E2]," [E2] records the 2018 recommendation and retrieval adds the 2023 update [E9]. UNITALIGN writes "Trastuzumab, pertuzumab, and taxane chemotherapy is the current first-line standard per 2023 ASCO guidelines [E2][E9]." The unit passes at $\tau = 0.72$, while surrounding later-line and contraindication text remains unchanged.