Where Does RAG Fail? A Triadic Alignment Benchmark for Diagnosing Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has become a dominant paradigm for knowledge-grounded text generation. Despite its prevalence, current evaluation approaches mostly score the generated response using either a single aggregate metric or multiple separately defined dimensions, without tracing failures to specific stages of the RAG pipeline. We propose TRIALIGN, a diagnostic benchmark that uses query-anchored triadic alignment to identify where grounded QA systems fail within a fixed evidence pool: in selecting sufficient evidence, covering required information, or attributing generated claims to the correct sources. Each query is decomposed into required information units (RIUs), enabling unit-level evaluation and a cascading failure attribution protocol that identifies whether each non-verified unit results from evidence selection, generation, or citation errors. TRIALIGN comprises 887 annotated queries, 3,298 required information units, and 8,575 unit–passage support links, with each query associated with a 100-passage candidate pool. Experiments with eight evidence selection strategies and four generators show that systems achieving similar answer completeness can exhibit substantially different failure profiles, with some being bottlenecked by evidence selection and others by generation, illustrating that aggregate scores alone can mask critical distinctions in system behavior. The code is anonymously available at https://anonymous.4open.science/r/TRIALIGN-FC11.

1 Introduction

The Retrieval-Augmented Generation (RAG) paradigm (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Gao et al., 2024) enables large language models to incorporate external evidence when answering information-seeking queries. As RAG systems are increasingly deployed in high-stakes domains such as biomedicine, legal analysis, and technical support, rigorous evaluation becomes essential. Current evaluation practice remains predominantly coarse-grained: aggregate metrics such as answer-level F1, faithfulness scores, or end-to-end accuracy indicate overall performance but provide little insight into why a system fails. Two systems with similar answer-level scores may fail for fundamentally different reasons: one because relevant evidence was not included in the selected context, and another because the generator failed to use evidence that was available.

Several lines of work have sought to move beyond monolithic evaluation: dimensional scoring (Es et al., 2024; Saad-Falcon et al., 2024), citation verification (Gao et al., 2023; Xu et al., 2025b; Li et al., 2024; Xu et al., 2025a), response-anchored diagnostics (Ru et al., 2024; Friel et al., 2025), and sub-question coverage (Xie et al., 2025). Each captures important but partial aspects: dimensional frameworks do not trace failures to specific stages; citation benchmarks ignore coverage of query needs; diagnostic tools rely on reference-answer recall rather than query-anchored unit tracing; sub-question methods do not evaluate evidence sufficiency. None jointly provides query-anchored units, explicit unit–evidence mappings, and stage-level failure attribution (see Appendix A for extended discussion).

We propose TRIALIGN, a query-anchored diagnostic benchmark that models grounded QA evaluation as triadic alignment. Rather than evaluating the response in isolation, TRIALIGN anchors evaluation in the query's information needs. Each query is decomposed into required information units (RIUs), the atomic information needs that a complete answer must satisfy. Each unit is evaluated along three alignment edges: whether selected evidence supports it, whether the response covers it, and whether cited passages support the corresponding claim. This triadic structure yields a unit-level diagnosis for each non-verified RIU, clas-

sifying the failure as evidence insufficiency, answer under-coverage, or attribution misuse. We instantiate TRIALIGN by augmenting ASQA (Stelmakh et al., 2022) with RIUs and unit-level support mappings, yielding 887 annotated queries with 3,298 RIUs and 8,575 unit–passage support links over 100-passage candidate pools. The annotations are produced through independent labeling and reconciliation, with quality-control details in §3.

Our contributions are threefold. **(1)** We formalize grounded QA evaluation as query-anchored triadic alignment with a cascading failure attribution protocol. **(2)** We release TRIALIGN: 887 queries, 3,298 RIUs, and 8,575 unit–passage support links over fixed 100-passage pools. **(3)** A factorial study of eight selectors and four generators shows that systems with similar aggregate scores often require different improvement strategies, a distinction only unit-level evaluation exposes.

2 TRIALIGN Framework

2.1 Task Formulation

Let Q denote a user query and $E(Q) = \{e_1, \ldots, e_m\}$ denote its candidate evidence pool. A RAG system selects a subset $S_k \subseteq E(Q)$ of k snippets and generates a response R with inline citations C, where each citation associates a response span with one or more passages in S_k.

We decompose each query into a set of *required information units* $U(Q) = \{u_1, \ldots, u_n\}$. A required information unit (RIU) is an atomic, self-contained information need that contributes to a complete answer. The unit set $U(Q)$ is constructed according to three desiderata: (1) *Atomicity*: each unit captures one distinct facet of the query and cannot be split into multiple independent information needs. (2) *Non-overlap*: different units address different aspects of Q, and satisfying one unit does not entail satisfying another. (3) *Query-anchoredness*: units are anchored in the query intent and its disambiguation structure, and are constructed independently of retrieved passages and system outputs. Figure 1 illustrates the RIU construction principles.

2.2 Triadic Alignment

Given $U(Q)$, selected evidence S_k, and response R, TRIALIGN evaluates each unit along three alignment edges. Evidence sufficiency checks whether the selected evidence supports the unit:

$$\text{supp}(u_i, S_k) = \mathbb{1}[\exists e_j \in S_k : e_j \vdash u_i]. \quad (1)$$

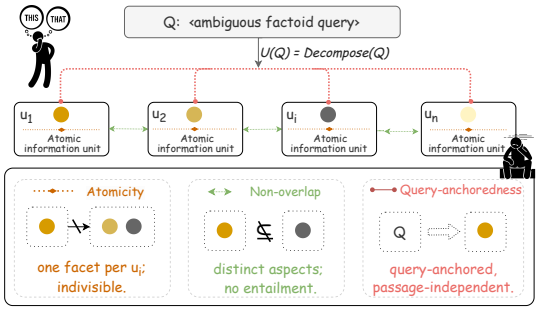

Figure 1: Illustration of RIU construction principles.

Here, $\mathbb{1}[\cdot]$ is 1 if the condition inside is true, 0 otherwise; $\text{supp}(u_i, S_k) = 1$ means at least one selected snippet supports unit u_i.

Answer coverage checks whether the response addresses the unit:

$$\text{cover}(u_i, R) = \mathbb{1}[R \text{ addresses } u_i]. \quad (2)$$

This indicator equals 1 if the generated response R contains information that satisfies unit u_i, independent of whether a citation is provided. Otherwise, it equals 0.

Attribution correctness checks whether the response's claim for a covered unit is linked to a supporting passage via a citation:

$$\text{attr}(u_i, R, S_k) = \mathbb{1}[\exists c_j \in C(u_i), \exists e \in S_k : c_j \mapsto e \wedge e \vdash u_i]. \quad (3)$$

Here $C(u_i) \subseteq C$ denotes the citations attached to the response segment addressing u_i, $c_j \mapsto e$ means citation c_j references passage e, and $e \vdash u_i$ means e supports unit u_i. The indicator equals 1 when the claim addressing u_i is correctly cited to supporting evidence; otherwise 0. Figure 2 illustrates the complete triadic alignment framework.

2.3 Metrics

Each alignment edge yields one or more metrics. For evidence sufficiency ($U \leftrightarrow S$), unit recall measures the fraction of query units covered by the selected evidence:

$$\text{U-Rec@}k = \frac{1}{|U|} \sum_{i=1}^{|U|} \text{supp}(u_i, S_k). \quad (4)$$

Unit precision measures the fraction of selected snippets that support at least one unit:

$$\text{U-Prec@}k = \frac{1}{|S_k|} \sum_{e_j \in S_k} \mathbb{1}[\exists u_i \in U : e_j \vdash u_i]. \quad (5)$$

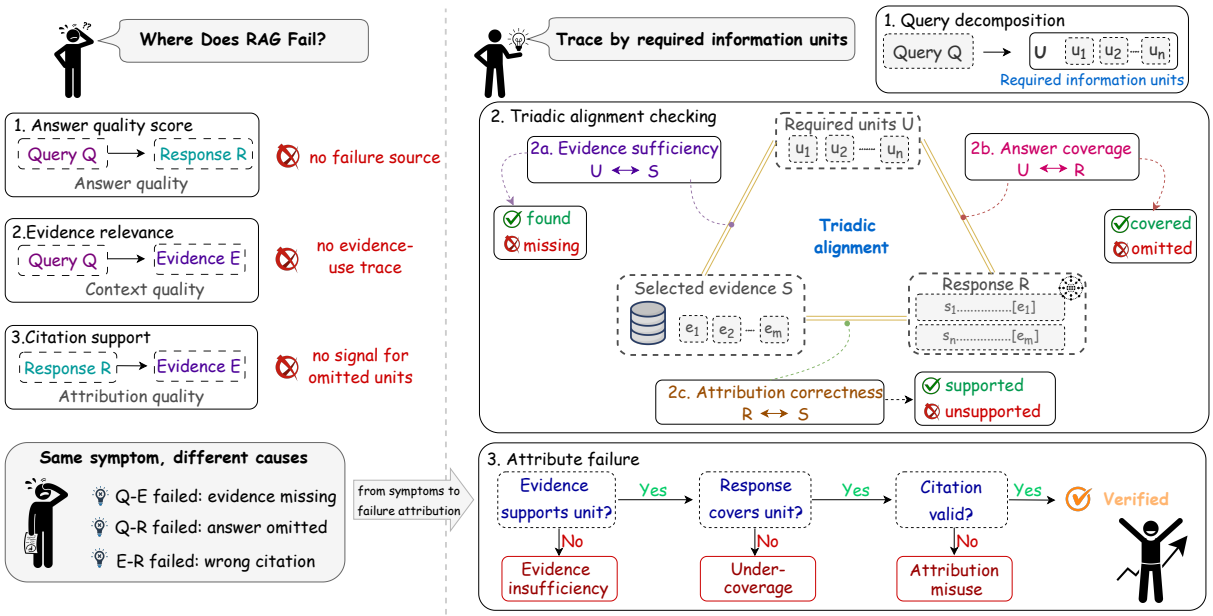

Figure 2: Overview of TRIALIGN. Required information units serve as query-anchored evaluation anchors across three alignment edges: evidence sufficiency ($U \leftrightarrow S$), answer coverage ($U \leftrightarrow R$), and attribution correctness ($R \leftrightarrow S$). Each non-verified unit is attributed to the earliest failed stage in the selection–generation–attribution pipeline.

For answer coverage ($U \leftrightarrow R$), answer completeness (AC) measures the fraction of query units addressed by the response:

$$\text{AC} = \frac{1}{|U|} \sum_{i=1}^{|U|} \text{cover}(u_i, R). \tag{6}$$

For attribution correctness ($R \leftrightarrow S$), citation recall measures the fraction of covered units that are correctly cited:

$$\text{Cite-Rec} = \frac{1}{|U_{\text{cov}}|} \sum_{u_i \in U_{\text{cov}}} \text{attr}(u_i, R, S_k) \tag{7}$$

where $U_{\text{cov}} = \{u_i \in U : \text{cover}(u_i, R) = 1\}$. Citation precision measures the fraction of citation instances whose referenced passage supports the associated claim:

$$\text{Cite-Prec} = \frac{1}{|C|} \sum_{c_j \in C} \mathbb{1}[\text{ref}(c_j) \vdash \text{claim}(c_j)] \tag{8}$$

where $\text{ref}(c_j)$ is the passage cited by c_j and $\text{claim}(c_j)$ is the response claim associated with c_j. Each citation marker references exactly one passage (i.e., `[N]` maps to a single $e_N \in S_k$); when multiple passages are cited for the same claim, each constitutes a separate citation instance. For macro averaging, queries with no covered RIUs are excluded from Cite-Rec computation, and queries generating no citations receive Cite-Prec = 0.

2.4 Failure Attribution Protocol

The alignment decisions above allow each non-verified unit to be assigned to the earliest failed stage in the RAG pipeline. We use a cascading protocol following the order evidence selection → generation → attribution:

- **Evidence insufficiency:** $\text{supp}(u_i, S_k) = 0$. The selected evidence does not support the unit.
- **Answer under-coverage:** $\text{supp}(u_i, S_k) = 1$ and $\text{cover}(u_i, R) = 0$. Supporting evidence was selected, but the response omits the unit.
- **Attribution misuse:** $\text{supp}(u_i, S_k) = 1$, $\text{cover}(u_i, R) = 1$, and $\text{attr}(u_i, R, S_k) = 0$. The response covers the unit with supporting evidence available, but the citation is missing or incorrect.

This ordering attributes each failure to the earliest responsible stage. Thus, if selected evidence does not support a unit, the failure is attributed to evidence insufficiency even if the response mentions the unit from parametric knowledge, since grounded QA requires claims to be verifiable against provided sources. A complete

Statistic	Value
Total queries	887
Avg. required units per query	3.72
Total required information units	3,298
Evidence pool size per query	100
Avg. supporting passages per unit	2.6

Table 1: Dataset statistics for TRIALIGN.

worked example of the diagnostic cascade is provided in Appendix G.

3 TRIALIGN Dataset

TRIALIGN uses queries from ASQA (Stelmakh et al., 2022), a long-form QA dataset of ambiguous factoid queries, paired with top-100 evidence pools from ALCE (Gao et al., 2023), retrieved by GTR-T5-XXL (Ni et al., 2022) from Wikipedia. Ambiguous factoid queries naturally require multi-faceted answers, making them appropriate for triadic evaluation; the fixed candidate pool further enables controlled comparison of evidence selection strategies.

Query Selection. We select queries whose information needs can be fully decomposed into atomic, non-overlapping units with verifiable evidence support. Specifically, we retain queries for which every identified information need has at least one supporting passage in the 100-passage pool. This yields a final set of 887 queries.

RIU Construction. For each retained query, annotators decompose the information needs into required information units following the three desiderata of §2.1 (atomicity, non-overlap, query-anchoredness). RIU construction was performed before unit–passage support labeling: annotators saw only the query text and ASQA disambiguation seeds, not the passage pool or system outputs. Pool-support filtering was applied post-hoc to ensure evaluability and did not inform RIU design. ASQA's disambiguation structure provides initial interpretations as a starting point, but the final RIU set is determined through independent annotation: annotators split compound needs into atomic units, merge near-duplicate interpretations, and discard items that fail the desiderata. Starting from 2,270 seed interpretations, only 68% are retained as-is; the remainder undergo splitting (452 compound seeds split into 1,710 resulting RIUs), merging, or removal, producing 3,298 final RIUs (a 45% net increase).

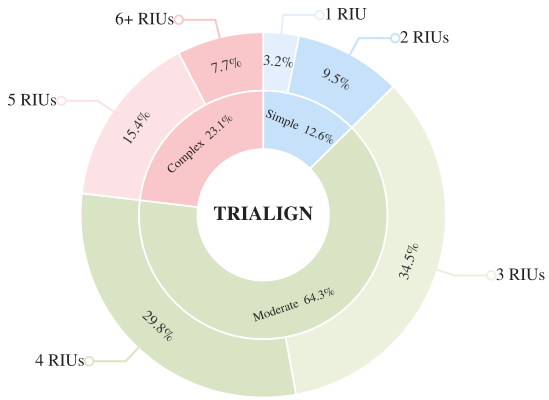

Figure 3: TRIALIGN query composition by RIU count.

Unit–Passage Support Annotation. The core annotation of TRIALIGN is the unit–passage support mapping: for each RIU, annotators identify which passages in the 100-passage pool provide sufficient evidence to satisfy it. To manage annotation cost over the large pool, we use a two-stage process: an initial candidate set is identified via provenance metadata, and annotators then verify these candidates and inspect additional passages as needed (details in Appendix B.4). This mapping connects query-level requirements to specific evidence, allowing the framework to determine whether a failure originates from evidence selection, generation, or citation. Two annotators independently label each unit–passage pair, with disagreements resolved by a senior annotator. The resulting matrix contains 8,575 binary support links across 3,298 units. Detailed annotation guidelines and quality-control procedures are provided in Appendix B.

Dataset Statistics. The resulting benchmark spans queries from simple cases with a single RIU to complex ones requiring six or more, with an average of 3.72 RIUs per query. Table 1 summarizes the headline statistics, and Figure 3 visualizes the RIU-count distribution. We classify queries into three complexity tiers based on RIU count: queries with 1–2 RIUs are Simple, those with 3–4 RIUs are Moderate, and those with 5 or more are Complex; the Moderate tier dominates the benchmark, accounting for nearly two-thirds of all queries. Each RIU is supported on average by 2.6 passages, and detailed breakdowns are reported in Appendix E.

System	Evidence		GPT-4o			Qwen2.5-72B			Llama-3.1-70B			Llama-3.1-8B		
	U-Rec	U-Prec	AC	C-Rec	C-Prec	AC	C-Rec	C-Prec	AC	C-Rec	C-Prec	AC	C-Rec	C-Prec
Random-5	**14.9**	6.3	**47.5**	23.6	62.2	44.2	19.8	57.2	41.8	17.4	54.1	34.2	10.8	43.5
BM25	37.6	20.6	52.2	56.2	57.8	48.5	47.5	53.2	46.1	42.8	50.3	38.6	28.3	40.5
DPR	55.9	31.5	63.0	74.5	67.0	58.1	62.4	61.6	55.2	56.1	58.3	44.8	37.2	46.9
Reranker	59.8	33.2	65.2	78.4	68.2	59.5	64.8	62.4	56.5	58.4	59.0	45.8	38.6	47.5
Top-5 (GTR)	63.8	39.5	**67.4**	**82.5**	56.4	62.0	70.6	51.9	58.6	63.7	49.1	**47.3**	**42.5**	39.5
MMR	70.2	35.8	69.1	87.3	55.8	63.5	74.8	51.3	60.1	67.5	48.5	48.7	44.9	39.1
LLM-Sel.	76.4	38.6	70.5	91.8	54.5	64.8	78.6	50.1	61.5	71.2	47.4	49.6	47.3	38.2
Oracle	**85.9**	31.9	**71.8**	95.1	53.7	65.9	81.5	49.4	63.5	77.2	46.7	53.1	55.8	37.6

Table 2: Main results on TRIALIGN ($k = 5$, all values %). U-Rec@5 and U-Prec@5 depend only on evidence selection (shared across generators). AC, Cite-Rec (C-Rec), and Cite-Prec (C-Prec) vary with generator capability. Oracle denotes a near-optimal greedy gold selector that uses the annotated support matrix to maximize unit-level evidence coverage within the fixed candidate pool (see footnote in §4.1).

Automatic Evaluation. Finally, while the RIU and support mappings provide a gold standard for evaluating evidence selection, scalable assessment of system responses requires automated judgments for answer coverage and citation correctness. We use Claude Opus 4.7 (Anthropic, 2026) as the primary automatic judge and validate its reliability through cross-family agreement with Gemini 3.1 Pro (Google DeepMind, 2026) and human evaluation. Details, including prompts and calibration results, are provided in Appendix C, with generation prompt templates in Appendix C.5. This setup supports reproducible evaluation at scale, with all judgments tied to individual unit–passage annotations.

4 Experiments

We conduct a factorial evaluation crossing eight evidence selection strategies with four generators, yielding 32 configurations. The goal is twofold: to quantify where systems fail across the triadic alignment edges, and to demonstrate that similar aggregate scores can mask fundamentally different failure profiles.

4.1 Experimental Setup

Evidence selection strategies. All systems select $k = 5$ passages from the same 100-passage GTR-T5-XXL candidate pool for each query. We choose $k = 5$ as a representative operating point: it is widely used in RAG practice and balances evidence coverage against context length; we ablate $k \in \{3, 5, 10\}$ in §4.5. We evaluate eight strategies covering different paradigms: **Random-5** (uniform sampling); **BM25** (lexical re-ranking (Robertson and Zaragoza, 2009)); **DPR** (dense re-ranking (Karpukhin et al., 2020)); **Reranker** (MiniLM cross-encoder (Wang et al., 2020)); **Top-5 (GTR)** (original GTR ranking); **MMR** ($\lambda = 0.7$ for diversity (Carbonell and Goldstein, 1998)); **LLM-Selector** (GPT-4o-based passage selection);[1] and **Oracle** (greedy set cover over the gold support matrix).[2]

Generators. We pair each selection strategy with four generators spanning different capability levels: GPT-4o (OpenAI, 2024), Qwen2.5-72B-Instruct (Qwen Team, 2024), Llama-3.1-70B-Instruct (Grattafiori et al., 2024), and Llama-3.1-8B-Instruct. The two Llama models isolate the effect of model size within a single family.

Implementation details. All generators receive the same structured prompt with numbered passages and an instruction to answer with inline citations in [N] format. All models use greedy decoding (temperature = 0) for reproducibility. GPT-4o is accessed via the OpenAI API; open-source models are served via vLLM (Kwon et al., 2023) on 8×A800-80GB GPUs. Full hyperparameters for all selection strategies are in Appendix D.

Evaluation. We report all five TRIALIGN metrics (§2.3), macro-averaged across queries. Results are deterministic under greedy decoding; for Random-5, we average over three independent samples.[3] Answer coverage and citation judgments use

[1] LLM-Selector uses the same model as the frontier generator. We discuss implications of this overlap in Limitations.

[2] Oracle greedily selects the passage covering the most uncovered RIUs at each step (greedy set cover). On a random 10% subset, exact ILP-optimal selection achieves at most 0.3% higher U-Rec, confirming that the greedy approximation is near-optimal in practice. It achieves U-Rec = 85.9% rather than 100% because some queries require more than $k = 5$ passages to cover all RIUs. When all RIUs are satisfied before k slots are filled (common for queries with few RIUs), the remaining slots are filled by the highest-ranked GTR passages, explaining why Oracle's U-Prec can be lower than Top-5's.

[3] Random-5 standard deviations across three samples are below 0.8 points for all metrics, confirming stability.

	GPT-4o			Qwen2.5-72B			Llama-3.1-70B			Llama-3.1-8B		
System	**Evid**	**Covr**	**Attr**	**Evid**	**Covr**	**Attr**	**Evid**	**Covr**	**Attr**	**Evid**	**Covr**	**Attr**
Random-5	**95.8**	3.9	0.3	93.3	4.8	1.9	91.8	5.4	2.8	88.4	6.5	5.1
BM25	**88.3**	10.6	1.1	81.1	13.4	5.5	77.7	14.5	7.8	70.1	16.4	13.5
DPR	83.1	15.2	1.7	69.2	21.6	9.2	63.9	23.1	13.0	52.9	25.4	21.7
Reranker	82.2	15.8	2.0	65.4	23.9	10.7	60.0	25.2	14.8	48.8	27.1	24.1
Top-5 (GTR)	81.5	16.5	2.0	64.3	24.6	11.1	57.8	26.6	15.6	**45.3**	**29.0**	**25.7**
MMR	75.1	21.9	3.0	56.8	29.4	13.8	50.1	31.0	18.9	38.1	32.2	29.7
LLM-Sel.	66.9	28.8	4.3	48.1	34.8	17.1	42.0	35.4	22.6	30.8	35.3	33.9
Oracle	44.5	47.7	**7.8**	30.5	45.9	**23.6**	27.7	43.4	**28.9**	**20.0**	**39.6**	**40.4**

Table 3: Failure attribution (% of failed RIU-units, $k = 5$). Evid = evidence insufficiency, Covr = answer under-coverage, Attr = attribution misuse. Each failed unit is classified into exactly one category via the diagnostic cascade (§2.4). The absolute count of evidence-insufficient units is identical across generators within each row (selection-determined); percentage differences arise because weaker generators produce more total failures (coverage + attribution), making evidence failures a smaller share of the larger failure pool.

Claude Opus 4.7 as the automatic judge (§4.6), which is independent of all evaluated generators.

4.2 Main Results

Table 2 presents the main results. We highlight five key findings.

Diminishing returns from evidence selection. U-Rec@5 spans 71 points from Random-5 (14.9) to Oracle (85.9), yet AC spans only 24 points for the strongest generator (47.5 to 71.8). The diminishing pattern is especially visible at the high end: LLM-Selector achieves U-Rec = 76.4 (89% of Oracle's coverage), yet its AC advantage over Top-5 is only 3.1 points (70.5 vs. 67.4). A generation ceiling exists independent of selection quality.

Selector refinement saturates before generator capacity. Holding the generator fixed, upgrading Top-5 to Oracle adds only 4.4 AC points for GPT-4o and 5.8 for Llama-3.1-8B. In contrast, holding selection at Top-5, upgrading Llama-3.1-8B to GPT-4o adds 20.1 AC points and 40.0 Cite-Rec points. This asymmetry is limited to the strong-selection regime; Random-5 to Top-5 still gains 19.9 AC points for GPT-4o.

Citation recall as a capability discriminator. Under Top-5 selection, Cite-Rec spans 40 points across generators (GPT-4o 82.5 → 8B 42.5), while AC spans only 20 points (67.4 → 47.3). Cite-Rec is substantially more sensitive to generator capability than AC in this setting, suggesting that correct attribution requires more than comprehension alone.

Citation precision peaks at mid-range selection. DPR and Reranker achieve the highest Cite-Prec (67–68%), while stronger selection strategies (Top-5, MMR, Oracle) show lower Cite-Prec (54–56%). One explanation is that when selected passages have broader or more overlapping coverage, the generator must disambiguate among several potentially supporting passages per claim, increasing citation mapping difficulty. Random-5 shows moderately high Cite-Prec (62% for GPT-4o) despite weak selection, but this is a small-denominator artifact: the model produces far fewer citations and restricts them to unambiguous matches.

Aggregate metrics mask failure diversity. Traditional text-overlap metrics (ROUGE-L, BERTScore) (Lin, 2004; Zhang et al., 2020) largely preserve the system ranking (Kendall $\tau \geq 0.86$; Appendix F.4) but provide no diagnostic information about failure sources. In particular, they cannot distinguish the two AC ≈ 47 configurations (Random-5+GPT-4o vs. Top-5+8B) whose failure profiles diverge sharply (Table 3).

4.3 Failure Attribution Analysis

Table 3 presents failure attribution across all configurations.

Same AC, opposite prescriptions. Consider two system pairs with near-identical answer completeness but divergent failure profiles:

- **Random-5 + GPT-4o (AC = 47.5) vs. Top-5 + 8B (AC = 47.3).** These systems differ by 0.2 points in AC, indistinguishable by any aggregate metric (bootstrap 95% CI for the difference: [−2.1, +2.5], $n = 887$). Yet 95.8% of Random+GPT-4o's failures are evidence insufficiency (improve evidence selection), while Top-

5+8B shows 45.3% evidence, 29.0% coverage, and 25.7% attribution failures (upgrade the generator).

- **Oracle + 8B (AC = 53.1) vs. BM25 + GPT-4o (AC = 52.2).** AC differs by 0.9 points (bootstrap 95% CI: [−1.8, +3.6]), yet the failure profiles are reversed: 88.3% of BM25+GPT-4o's failures are evidence insufficiency (selection is the bottleneck), while Oracle+8B shows only 20.0% evidence, with 39.6% coverage and 40.4% attribution failures (the generator is the bottleneck).

Selection upgrades expose different downstream bottlenecks. The Random→Oracle progression shows how improved selection shifts failure composition, but the shift depends on the generator:

- **Fixed GPT-4o:** Coverage failure share rises from 3.9% to 47.7%; once selection is strong, the generation ceiling becomes the dominant bottleneck. MMR and LLM-Selector occupy intermediate positions (21.9% and 28.8%), showing that diversity-based and LLM-guided selection progressively shift the failure profile away from evidence toward generation.

- **Fixed 8B:** Attribution rises from 5.1% to 40.4% as the weak model's inability to produce correct citations surfaces once evidence is available.

The same selection upgrade produces different "next bottleneck" diagnoses depending on generator capability, a distinction only visible through per-unit failure attribution.

Attribution misuse and model scale. Under Oracle conditions, the share of failed units attributed to citation misuse ranges from GPT-4o 7.8% to Qwen 23.6%, 70B 28.9%, and 8B 40.4%. The smallest model's attribution share is roughly 5× higher than GPT-4o's. Qwen2.5-72B achieves a lower attribution share than Llama-70B at similar parameter count (23.6% vs. 28.9%), indicating that citation capability depends on training methodology as well as scale. To quantify the practical impact, we define *Verified-AC*—the fraction of RIUs both covered and correctly attributed (Appendix F.5). Under Oracle, Verified-AC drops from 69.4% (GPT-4o) to 21.3% (8B), compared to AC of 71.8% vs. 53.1%—meaning GPT-4o loses only 2.4 pp from attribution failures while Llama-8B loses 31.8 pp, revealing that weaker models cover content but largely fail to ground it with correct citations. We further analyze per-complexity failure distributions (Appendix F.1) and context size effects (Appendix F.3).

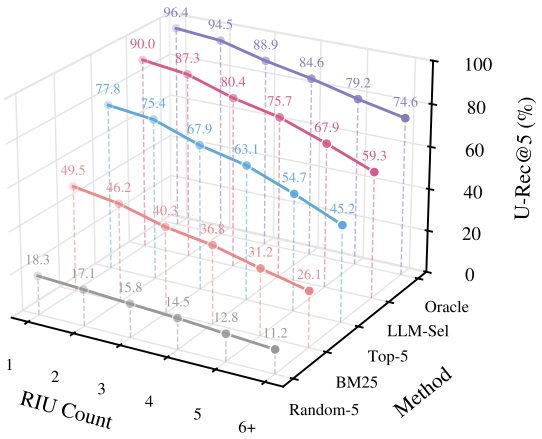

Figure 4: U-Rec@5 (%) by query complexity. U-Rec depends only on evidence selection, not on the generator. More complex queries are harder to satisfy with a fixed $k = 5$ budget.

Unsupported coverage. Our triadic design also reveals when generators produce answers without evidence support: units with supp $= 0$ yet cover $= 1$ are addressed without any selected evidence backing. Under Random-5, GPT-4o covers 42.3% of evidence-insufficient units, whereas Llama-3.1-8B covers only 18.4%; even under Top-5, GPT-4o still covers 30.1%. This asymmetry explains why strong generators maintain high AC despite weak evidence selection, and highlights that only the triadic framework can distinguish genuinely grounded answers from those relying on parametric knowledge. Full results across all configurations are in Appendix F.2.

4.4 Per-Complexity Analysis

The preceding analysis aggregates across all queries. We next ask whether query complexity, measured by the number of RIUs, modulates these findings.

Figure 4 shows that U-Rec@5 decreases monotonically with query complexity across all strategies. For queries with 6+ RIUs, even Top-5 achieves only 45.2% coverage, and a fixed $k = 5$ budget becomes increasingly inadequate as needs broaden. The gap between Top-5 and Oracle widens with complexity (18.6 points for 1-RIU queries vs. 29.4 for 6+-RIU queries), indicating greater headroom for improvement on complex

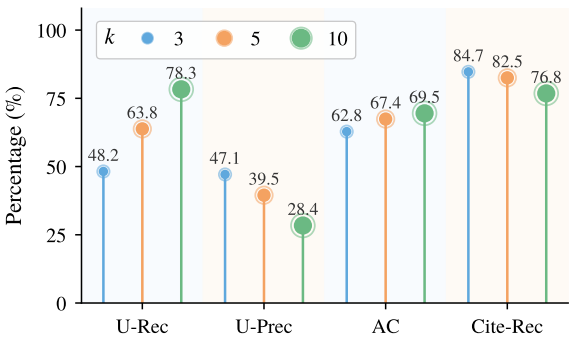

Figure 5: Effect of passage count k on TRIALIGN metrics (GPT-4o, Top-5 selection, all %).

queries but also that even oracle selection cannot fully close the gap under a fixed k budget. Systems targeting complex queries may benefit from diversity-based selection or adaptive k allocation. Failure attribution by complexity (Appendix F.1) confirms this pattern: evidence insufficiency accounts for 76.3% of failures on 1-RIU queries but 92.4% on complex queries (6+ RIUs), leaving little room for coverage or attribution failures to manifest.

4.5 Context Size Ablation

The complexity analysis above suggests that a fixed $k = 5$ budget constrains coverage for complex queries. We now vary k to examine the trade-offs of expanding context.

Figure 5 reveals a coverage–attribution trade-off. Increasing k from 3 to 10 improves U-Rec by 30 points and AC by 6.7 points, but Cite-Rec *decreases* from 84.7 to 76.8. One likely factor is increased citation mapping difficulty: with more candidate passages, the generator must identify the correct source among a larger set. A compositional effect may also contribute, since higher k covers additional (potentially harder) units whose citations are more challenging. Regardless of mechanism, the empirical trade-off implies that practitioners must balance coverage gains against attribution degradation when choosing context size.

4.6 Judge Calibration

We validate the automatic judge through two complementary approaches (Table 4). For cross-family agreement, we sample 200 RIU–response pairs for coverage and 200 claim–citation pairs for attribution, stratified by RIU count and generator (positive rate: 64% coverage, 71% attribution); Claude Opus 4.7 and Gemini 3.1 Pro agree on >89% of judgments ($\kappa = 0.82$ coverage, 0.78 attribution). For judge-vs-human validation, two annotators independently label 100 examples per task (positive rate: 58% coverage, 65% attribution; raw inter-annotator $\kappa = 0.81$), with disagreements adjudicated by the senior annotator. Opus 4.7 achieves F1 > 86 and $\kappa \geq 0.79$ against these gold labels. Evidence sufficiency is computed directly from the gold unit–passage support matrix, requiring no judge.

Judgment Task	**Agree.(%)**	**F1**	κ
Cross-family (Opus 4.7 vs. Gemini 3.1 Pro, n = 200):			
Answer coverage	91.2	–	0.82
Citation correctness	89.5	–	0.78
Judge vs. human gold (n = 100, Opus 4.7 output) against adjudicated labels from 2 annotators):			
Answer coverage	–	88.6	0.84
Citation correctness	–	86.3	0.79

Table 4: Automatic judge reliability. Cross-family: agreement between two LLM judges (Opus 4.7 vs. Gemini 3.1 Pro). Judge vs. human gold: F1 and κ of Opus 4.7 judgments against human-adjudicated labels (two annotators label independently, senior annotator resolves disagreements; raw inter-annotator $\kappa = 0.81$ for both tasks). Evidence sufficiency is computed from gold annotations and does not require a judge.

5 Conclusion

We have introduced TRIALIGN, a query-anchored diagnostic benchmark that evaluates RAG through triadic alignment among query requirements, selected evidence, and cited responses. By attributing each non-verified RIU to evidence insufficiency, answer under-coverage, or attribution misuse, TRIALIGN turns aggregate scores into diagnostic profiles indicating specific improvement targets. Our factorial evaluation of eight evidence selection strategies and four generators shows that similar answer-level performance can mask divergent failure sources. The same selection improvement can expose different downstream bottlenecks depending on the generator, and citation recall is much more sensitive to generator capability than answer completeness. These findings illustrate the practical value of unit-level diagnosis: without it, developers risk optimizing the wrong pipeline component. We release TRIALIGN to support more targeted evaluation and development of RAG systems. Future work includes extensions to multi-turn and cross-lingual settings.

Limitations

TRIALIGN focuses on 887 English queries and 3,298 required information units derived from ASQA's ambiguous factoid questions. While sufficient for evaluating ambiguous factoid QA, it does not cover other query types (multi-hop, procedural, or opinion-seeking) or cross-lingual settings, each of which raises distinct retrieval and annotation challenges that we leave to future work.

The evidence pools consist of pre-retrieved top-100 passages from GTR-T5-XXL rather than live retrieval over the full 21M-passage Wikipedia corpus. This controlled setup enables precise unit-level annotation and fair cross-system comparison, but may underestimate the difficulty of fully open-domain retrieval, where iterative or agentic strategies could yield different results. Furthermore, because TRIALIGN retains only queries whose RIUs are all supported within the fixed top-100 pool, it does not measure failures where the candidate pool itself lacks required evidence. This filtering improves diagnostic controllability but may underrepresent hard open-domain cases where no relevant passage is retrievable.

Automated evaluation relies on Claude Opus 4.7, which may introduce model-specific biases. Although cross-model validation with Gemini 3.1 Pro shows >89% agreement and human validation yields $\kappa \geq 0.79$, reliability may still vary on edge cases or specialized domains underrepresented in the judge's training distribution. Separately, the LLM-Selector and the frontier generator both use GPT-4o, which could introduce a shared-model bias; cross-generator experiments (pairing the LLM-Selector with non-GPT generators) provide partial control for this dependency.

Ethics Statement

TRIALIGN is built on publicly available data: queries from ASQA (Stelmakh et al., 2022) and Wikipedia passages from ALCE (Gao et al., 2023). We will release the TRIALIGN annotations, evaluation scripts, and system outputs on GitHub with documentation and license information. No private or sensitive information is collected or generated. All annotators were graduate students compensated at USD 18/hour (above local minimum wage), informed of the research purpose and public release plan, and given the right to withdraw. The benchmark evaluates existing model outputs and does not involve deceptive interaction or personal data collection. Automated evaluation using Claude Opus 4.7 and Gemini 3.1 Pro was conducted via official APIs under standard terms of service. We acknowledge that LLM-based judges may exhibit biases inherited from training; our cross-family and human validation partially mitigates but does not eliminate this risk.

References

Anthropic. 2026. Claude opus 4.7 system card. Anthropic system cards. Model system card for Claude Opus 4.7.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. 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*, pages 150–158.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Google DeepMind. 2026. Gemini 3.1 pro model card. Google DeepMind model card. Accessed: 2026-05-18.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Albert Sacks, Alex Letman, Alex Lumber, Richard Tanburn, Richard Thomas, and Robert Thompson. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *ICML*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. Attributionbench: How hard is automatic attribution evaluation? *Preprint*, arXiv:2402.15089.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Sacha Muller, Antonio Loison, Bilel Omrani, and Gautier Viaud. 2025. GroUSE: A benchmark to evaluate evaluators in grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4510–4534.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. MIRAGE: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Preprint*, arXiv:2408.08067.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. 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*, pages 233–256.

Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. Garage: A benchmark with grounding annotations for rag evaluation. *Preprint*, arXiv:2506.07671.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Kaige Xie, Philippe Laban, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Do RAG systems cover what matters? evaluating and optimizing responses with sub-question coverage. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5836–5849.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025a. ALiiCE: Evaluating positional fine-grained citation generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 545–561.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. 2025b. CiteEval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32759–32778.

Jiajie Zhang, Juntao Shang, Zhuocheng Zhang, Junhao Liu, Juntao Huang, Tong Lu, Zhiyang Wu, Yuxiao Zhu, Kaizhi Leng, and Kaiwen Bi. 2024. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. *arXiv preprint arXiv:2409.02897*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

A Related Work

A.1 RAG Evaluation Frameworks and Diagnostics

Early frameworks for RAG evaluation focused on decomposing quality into independently assessed dimensions. Es et al. (2024) propose RAGAs, measuring context relevance, answer relevance, and faithfulness via LLM-based scoring, while Saad-Falcon et al. (2024) introduce ARES, which reduces annotation overhead using lightweight classifiers. Both frameworks provide practical, scalable evaluation but lack mechanisms to attribute failures to retrieval, generation, or citation stages.

Diagnostic frameworks attempt finer-grained analysis. RAGChecker (Ru et al., 2024) evaluates atomic claims against retrieved chunks and reference answers, producing module-level metrics. However, it is *response-anchored*: it evaluates claims present in the output and detects omissions only via claim recall against reference answers, without providing query-anchored omission detection or explicit unit–evidence tracing. Its diagnosis is correlational rather than stage-specific, and claim recall requires reference answers. Friel et al. (2025) introduce RAGBench, a large-scale, cross-domain evaluation with sentence-level support labels, but these remain binary and do not connect evidence to specific query needs. GroUSE (Muller et al., 2025) demonstrates that even strong LLM evaluators miss critical failure modes, motivating structured, unit-level protocols.

A.2 Benchmarks with Controlled Evidence

Several benchmarks provide fixed candidate pools to study retrieval–generation interactions. MIRAGE (Park et al., 2025) and GaRAGe (Sorodoc et al., 2025) offer curated passage pools with relevance annotations, enabling researchers to isolate the effects of retrieval quality. ALCE (Gao et al., 2023) further provides end-to-end citation QA evaluation. While informative, these benchmarks focus on aggregate metrics or passage-level relevance, and do not support explicit mapping between query-level requirements and evidence for stage-level failure attribution. TRIALIGN builds on ALCE's controlled setting but extends it with RIUs and unit-level evidence mappings across 887 queries, enabling a factorial study of eight evidence selection strategies and four generators for fine-grained diagnostic analysis.

A.3 Citation, Attribution, and Coverage Evaluation

Trustworthy generation depends on accurate citation and attribution. AttributionBench (Li et al., 2024), CiteEval (Xu et al., 2025b), ALiiCE (Xu et al., 2025a), and LongCite (Zhang et al., 2024) assess whether claims are correctly grounded, revealing the difficulty of automated verification even for strong LLMs. FActScore (Min et al., 2023) introduced atomic-claim decomposition for factuality evaluation, which informs claim-level assessment in our work. Sub-question methods (Xie et al., 2025) measure query–response coverage but do not jointly evaluate evidence sufficiency or provide unit-to-evidence mappings.

TRIALIGN unifies these perspectives under a query-anchored RIU framework, evaluating all three edges of the query–evidence–response triangle. This allows systematic attribution of failures to evidence selection, generation, or citation, which prior approaches evaluating edges separately cannot achieve.

Summary. As summarized in Table 5, prior work captures important but partial aspects of RAG evaluation, including dimension-wise scoring, controlled retrieval–generation comparison, query–response coverage, and citation attribution. However, none jointly provides query-anchored units, omission detection, explicit unit–evidence tracing, citation attribution, and stage-level diagnosis within a single protocol. TRIALIGN addresses this limitation by introducing RIUs as stable anchors across queries, evidence, and responses, enabling controlled, fine-grained diagnosis that guides targeted RAG improvements.

Prior work	Q-Units	Omission	U–E Trace	Citation	Stage Diag.	R–G Comp.
RAGAs (Es et al., 2024)	×	×	×	×	×	×
ARES (Saad-Falcon et al., 2024)	×	×	×	×	×	×
RAGChecker (Ru et al., 2024)	×	△	×	△	△	×
RAGBench (Friel et al., 2025)	×	×	×	△	×	△
MIRAGE (Park et al., 2025)	×	×	×	×	×	✓
GaRAGe (Sorodoc et al., 2025)	×	×	×	△	×	✓
ALCE (Gao et al., 2023)	×	×	×	✓	×	✓
AttributionBench (Li et al., 2024)	×	×	×	✓	×	×
CiteEval (Xu et al., 2025b)	×	×	×	✓	×	×
ALiiCE (Xu et al., 2025a)	×	×	×	✓	×	×
LongCite (Zhang et al., 2024)	×	×	×	✓	×	×
Sub-question coverage (Xie et al., 2025)	△	✓	×	×	×	×
TRIALIGN	✓	✓	✓	✓	✓	✓

Table 5: Comparison with prior RAG evaluation work. Q-Units denotes query-anchored information units defined independently of system responses. Omission denotes detection of missing query requirements. U–E Trace denotes explicit links between query-side units and supporting evidence. Citation denotes verification of response–source attribution. Stage Diag. denotes attribution of failures to evidence selection, generation, or citation stages. R–G Comp. denotes controlled comparison of retrieval and generation components under shared evidence settings. ✓ indicates explicit support, △ partial or indirect support, and × absence of the capability.

B Annotation Guidelines and Quality Control

We provide the complete annotation guidelines used during dataset construction. Annotators received a detailed instruction manual covering: (1) required information unit (RIU) construction with worked examples across ambiguity types and complexity levels; (2) evidence support judgment criteria with boundary cases; and (3) claim extraction and citation verification protocols used for judge validation.

B.1 Annotator Recruitment, Background, and Compensation

We recruited six annotators in total: four primary annotators responsible for RIU construction and unit–passage support labeling, one senior annotator (a graduate researcher in NLP with prior experience in QA dataset annotation) who served as the final adjudicator and quality reviewer, and one independent annotator who performed the held-out spot check described in §B.6. All annotators were graduate students with a computer science or computational linguistics background and self-reported English proficiency at C1 (CEFR) or above. Each annotator was compensated at an hourly rate of USD 18, exceeding the prevailing minimum wage in the recruitment region. Annotators were informed of the research purpose, the public release plan for the resulting benchmark, and their right to withdraw before signing the work agreement.

B.2 Annotator Training

All four primary annotators completed a pilot annotation stage before working on the final dataset. The pilot set contained 50 queries sampled from ASQA and was used to refine the annotation guidelines. During the pilot, annotators independently constructed RIUs and labeled unit–passage support links. Disagreements were discussed with the senior annotator and converted into boundary examples and decision rules in the final guideline. The pilot annotations were not included in the final benchmark statistics. Following the pilot, each annotator independently re-annotated a 30-query calibration batch; only annotators reaching pairwise Cohen's $\kappa \geq 0.75$ with the senior annotator's reference labels were retained for the main annotation.

B.3 Required Information Unit Construction

ASQA disambiguation sub-questions provide seed interpretations for each query. Annotators use these seeds as a starting point and construct the final RIU set by determining whether each seed is atomic, non-overlapping, and query-anchored. Annotators apply four operations during construction:

- **Retain.** A seed is retained as an RIU if it already expresses a distinct and necessary information need.
- **Split.** A seed is split if it combines multiple distinct information needs that require separate evidence (e.g., a sub-question asking about both causes and effects).

- **Merge.** Two seeds are merged if they refer to the same interpretation or ask for the same information using different wording.
- **Discard.** A seed is discarded if it is redundant, trivial, not required by the original query, or too vague to be verified against evidence.

The final RIU set for each query is produced through independent annotation by three of the four primary annotators followed by reconciliation. We compute Fleiss' κ (Fleiss, 1971) on the initial retain/split/merge/discard decisions (before reconciliation) to capture inter-annotator agreement.

Example. For the query *"What are the main causes, symptoms, and treatments of iron deficiency anemia?"*, annotators identify separate RIUs for dietary causes, medical causes, symptoms, dietary or supplementation-based treatments, and medical treatments for severe cases. This illustrates that RIUs should be specific enough to support binary evidence judgments while remaining broad enough to accommodate different valid phrasings.

B.4 Unit–Passage Support Annotation

For each query Q, annotators identify which passages in the 100-passage candidate pool support each RIU. A passage is marked as supporting an RIU if it contains enough information to answer the unit without relying on unstated assumptions. Support is judged at the unit level rather than the query level: a passage may be topically relevant to the query but still fail to support a particular RIU.

Because exhaustively labeling all $n \times 100$ pairs would be prohibitively expensive, we adopt a two-stage annotation strategy. First, we identify high-likelihood candidate passages using ASQA's `wikipages` provenance field, which links each gold answer to its source Wikipedia page(s); passages originating from these pages form the initial inspection set, typically 8–15 passages per query. Second, annotators verify these candidates and actively search the remaining pool for additional supporting passages—they are instructed to scan titles and opening sentences of all 100 passages and pull any that might support an uncovered RIU into the full inspection set. This two-stage design concentrates effort on high-yield candidates while still allowing discovery of supporting passages outside the provenance set. To estimate recall of this process, the independent spot-check annotator (described in §B.6) additionally performed an exhaustive scan of all 100 passages for a random 5% subset of queries; no additional supporting passages were found beyond those already in the reconciled matrix, suggesting that the two-stage strategy achieves near-complete coverage. Each pair (u_i, e_j) receives a binary support label.

Two of the primary annotators independently label each candidate pair, and disagreements are resolved by the senior annotator. We report Cohen's κ (Cohen, 1960) computed on the two independent annotations *before* adjudication, to reflect raw annotator agreement rather than post-reconciliation consistency. The reconciled labels form the unit–passage support matrix used to compute evidence sufficiency metrics.

B.5 Response-Side Validation Annotation

For judge validation, annotators label system-generated responses at the claim level. Each response is first decomposed into atomic claims following Min et al. (2023). Annotators then map each claim to the RIU it addresses, if any. Claims that do not correspond to any RIU are marked as extraneous.

For each claim with an inline citation, annotators verify whether the cited passage supports the claim. Each claim–citation relation is assigned one of four labels:

- **Correct citation:** the cited passage supports the claim.
- **Wrong citation:** the cited passage does not support the claim.
- **Missing citation:** the claim requires support but no citation is provided.
- **Redundant citation:** the citation is attached to a claim that does not require external support or does not address any RIU.

These annotations are used to validate automatic judgments for answer coverage and attribution correctness (Appendix C).

B.6 Quality Control

We employ four quality-control procedures.

Pilot refinement. The annotation guideline was refined through a 50-query pilot. Disagreements observed during the pilot were reviewed and converted into explicit decision rules.

Annotation Task	Agreement
RIU construction (3 annotators)	Fleiss' $\kappa = 0.79$
Unit–passage support (2 ann., pre-adj.)	Cohen's $\kappa = 0.81$
Claim mapping and citation (2 ann.)	Cohen's $\kappa = 0.81$
Support-label spot check (independent)	96.2% agreement

Table 6: Inter-annotator agreement for TRIALIGN annotations. Fleiss' κ (Fleiss, 1971) is used for multi-annotator tasks; Cohen's κ (Cohen, 1960) for pairwise tasks. Both κ values are computed on raw annotations *before* reconciliation.

Independent annotation and reconciliation. RIU construction, unit–passage support annotation, and response-side validation were each conducted independently by multiple annotators. Disagreements were resolved through discussion or by the senior annotator.

Senior annotator review. All reconciled RIU sets were reviewed by the senior annotator for granularity, non-overlap, and query-anchoredness. The senior annotator is a graduate researcher with prior experience in QA dataset annotation and is independent of the four primary annotators (i.e., does not contribute initial labels for reconciliation). Approximately 8% of unit sets received revisions during this stage; the most common revisions were further splitting of compound units and tightening of overly broad units.

Support-label spot check. An independent sixth annotator (not involved in the original labeling) re-annotated the unit–passage pairs for a random 10% subset of queries (89 queries, 331 RIUs, covering all candidate pairs inspected during the original annotation). Overall agreement with the final reconciled labels was 96.2%; broken down by label, agreement was 93.8% on positive support links and 97.1% on negative pairs, confirming that the high overall figure is not driven solely by the majority-negative class. In an additional exhaustive scan of all 100 passages for a random 5% subset (44 queries), no positive links were found beyond those already in the reconciled matrix. Table 6 summarizes inter-annotator agreement across all annotation stages.

C Automatic Judge and Prompt Templates

C.1 Overview

Human annotations provide gold RIU sets and unit–passage support mappings. Since all retrieval-augmented systems select evidence from the fixed 100-passage pool, evidence sufficiency is computed directly from the annotated unit–passage support matrix. For response-side evaluation, we use an automatic judge to assess answer coverage and attribution correctness at scale. Headline judge calibration numbers are reported in §4.6 (Table 4); this appendix provides the full prompt templates and a finer-grained breakdown by query complexity.

C.2 Answer Coverage Judging

For each RIU–response pair (u_i, R), the judge determines whether the response contains information that addresses the RIU. The judge is instructed to ignore citation correctness at this stage and focus only on whether the required information appears in the response.

Prompt template.

```
You are evaluating whether a response
addresses a specific information need.

Query: {query}
Required information unit: {unit}
Response: {response}

Does the response contain information
that addresses this required information
unit?  The information need not be
phrased identically. Answer YES or NO,
then explain briefly.
```

C.3 Attribution Correctness Judging

For each covered RIU, we identify the response claim addressing that unit and inspect its attached citations. A citation is correct if the cited passage provides sufficient support for the claim. If no citation is attached to a claim that requires evidence, the unit is marked as unattributed.

Prompt template.

```
You are verifying whether a citation
correctly supports its associated claim.

Claim: {claim_text}
Cited passage: {passage_text}
Required information unit: {unit}

Does the cited passage provide direct
textual support for this specific claim
with respect to the required information
unit? Answer YES or NO, then explain
briefly.
```

C.4 Per-Complexity Judge Calibration

Headline cross-family and human agreement results are reported in Table 4. Table 7 provides a complementary breakdown by query complexity, measured by the number of RIUs. Agreement decreases modestly with complexity but remains

Complexity	Coverage κ	Citation κ
1 RIU	0.89	0.85
2 RIUs	0.86	0.82
3 RIUs	0.84	0.80
4 RIUs	0.82	0.77
5 RIUs	0.79	0.75
6+ RIUs	0.76	0.72

Table 7: Per-complexity judge–human agreement (Cohen's κ, computed on the $n = 100$ human-validation subset of §4.6, stratified by RIU count).

Role	HuggingFace checkpoint
DPR question enc.	`facebook/dpr-question_encoder-single-nq-base`
DPR context enc.	`facebook/dpr-ctx_encoder-single-nq-base`
Cross-encoder	`cross-encoder/ms-marco-MiniLM-L-12-v2`

Table 8: Model checkpoints used by selection components in Appendix D.2.

substantial across all bins, indicating that the judge does not break down on more demanding queries.

C.5 Generation Prompt

All generators receive the same structured prompt consisting of the query, the selected passages, and an instruction to answer comprehensively with inline citations.

```
Answer the following question using
ONLY the provided passages. Cite
passages using [N] format inline. Be
comprehensive and address all aspects
of the question.

Question: {query}

Passages:
[1] {passage_1}
[2] {passage_2}
...
[k] {passage_k}

Answer:
```

D Implementation Details

This appendix lists implementation specifics not covered in §4.1, focusing on hyperparameters, hardware, and reproducibility settings.

D.1 Decoding and Generation Settings

All generators use greedy decoding (temperature = 0, top-p = 1) with a maximum of 1024 generated tokens. The same chat template is used for all open-source generators, with a system message instructing the model to ground answers in the provided passages and to use `[N]` inline citation format. Open-source models are served with vLLM v0.6.3 (Kwon et al., 2023) using tensor parallelism on 8×A800-80GB GPUs.

D.2 Evidence Selection Hyperparameters

All eight selection strategies operate over the same 100-passage candidate pool. We document hyperparameters and concrete implementations here; the specific HuggingFace checkpoints used by the dense and cross-encoder components are listed in Table 8. High-level descriptions of each strategy are in §4.1.

- **Random-5:** three independent samples are drawn with seeds 42, 123, 456; reported metrics are averages over the three runs.
- **BM25:** the `rank-bm25` Python package with default parameters $k_1 = 1.2$, $b = 0.75$, applied within the 100-passage pool.
- **DPR:** dual encoders (Table 8); re-ranking uses inner-product scores on L2-normalized embeddings.
- **Reranker:** a MiniLM cross-encoder (Table 8) (Wang et al., 2020), scoring each (query, passage) pair and taking the top 5.
- **Top-5 (GTR):** first five passages in the original GTR-T5-XXL (Ni et al., 2022) ranking provided by ALCE (Gao et al., 2023).
- **MMR:** GTR query–passage cosine similarity as the relevance term and inter-passage cosine similarity as the diversity penalty, with $\lambda = 0.7$; greedy selection until five passages are chosen.
- **LLM-Selector:** GPT-4o prompted with the query and the 100 candidate passages, each represented by its title and first 200 tokens. The prompt asks the model to select the five passage indices that best cover all aspects of the query; decoding uses temperature 0.
- **Oracle:** greedy set-cover over the gold unit–passage support matrix; at each step, the passage covering the most uncovered RIUs is added. When all RIUs are satisfied before five slots are filled, the remaining slots are filled by the highest-ranked GTR passages not already selected.

Operation	Seeds	Resulting RIUs
Retain as-is	1,550	1,550
Split	452	1,710
Merge	76	38
Discard	192	0
Total	**2,270**	**3,298**

Table 9: Construction flow from seed interpretations to final RIUs. "Seeds" counts the inputs to each operation; "Resulting RIUs" counts the units produced after the operation.

D.3 Compute Budget

End-to-end response generation across all (selection strategy, generator) configurations took approximately 410 GPU-hours on 8×A800-80GB. Judge evaluation cost approximately USD 240 in Claude Opus 4.7 API usage and USD 35 in Gemini 3.1 Pro usage for the cross-family validation subset.

E Additional Dataset Statistics

Table 9 reports the construction flow from seed interpretations to final RIUs. We start from 2,270 seed interpretations (derived from ASQA disambiguation sub-questions) across the 887 retained queries (an average of 2.56 seeds per query). After the construction operations of Appendix B.3, we obtain 3,298 final RIUs (an average of 3.72 per query). The net increase is driven primarily by splits: compound seeds tend to bundle multiple information needs, and splitting them is required to satisfy the atomicity desideratum.

Table 10 reports the per-bucket query distribution underlying the three complexity tiers introduced in §3. The distribution is right-skewed: 3-RIU and 4-RIU queries together cover 64.3% of the benchmark, while 1-RIU and 6+-RIU queries are relatively rare at 3.2% and 7.7%. The annotated unit–passage matrix contains 8,575 binary (u_i, e_j) support links in total, equivalent to roughly 2.6 supporting passages per RIU on average.

F Additional Analyses

F.1 Per-Complexity Failure Analysis

We analyze how failure attribution varies with query complexity, measured by the number of RIUs. Table 11 shows the breakdown for the strongest (GPT-4o) and weakest (Llama-8B) generators, averaged across the seven non-Oracle selection strategies. The 76.3%/92.4% figures cited in §4.4 correspond to the GPT-4o column of this table.

RIU count	Queries	Share
1	28	3.2%
2	84	9.5%
3	306	34.5%
4	264	29.8%
5	137	15.4%
6+	68	7.7%
Total	**887**	**100.0%**

Table 10: Per-bucket distribution of queries by RIU count. The three complexity tiers in §3 aggregate 1–2 RIUs as Simple, 3–4 RIUs as Moderate, and 5+ RIUs as Complex.

	GPT-4o			Llama-8B		
RIU Count	**Evid**	**Covr**	**Attr**	**Evid**	**Covr**	**Attr**
1	76.3	21.5	2.2	57.8	25.6	16.6
2	79.8	18.2	2.0	62.4	22.8	14.8
3	84.1	14.3	1.6	67.8	19.5	12.7
4	88.3	10.4	1.3	74.1	15.8	10.1
5	90.6	8.2	1.2	78.3	12.9	8.8
6+	92.4	6.6	1.0	80.5	11.6	7.9

Table 11: Failure attribution by query complexity, averaged across the seven non-Oracle selection strategies. All values are percentages of failed units.

For GPT-4o, evidence insufficiency rises from 76.3% to 92.4% as complexity increases, with coverage and attribution failures both shrinking in relative terms. For Llama-8B, the same trend holds for evidence insufficiency (57.8%→80.5%), but attribution misuse remains substantial even at the highest complexity (still 7.9%), reflecting the weaker model's persistent citation difficulty. At every complexity level, Llama-8B shows higher attribution failure than GPT-4o, consistent with the generator-capability effect in §4.3.

F.2 Unsupported Coverage

Table 12 reports the unsupported coverage rate—the fraction of evidence-insufficient units (supp = 0) that the generator nevertheless covers (cover = 1)—across all selection strategies and generators. This metric quantifies how much each generator relies on parametric knowledge or indirect inference rather than provided evidence.

Two patterns emerge. First, stronger generators consistently show higher unsupported coverage: GPT-4o covers 42.3% of unsupported units under Random-5, compared to 18.4% for Llama-8B. This explains why strong generators maintain

System	GPT-4o	Qwen-72B	70B	8B
Random-5	42.3	35.1	29.8	18.4
BM25	38.5	31.6	26.4	15.7
DPR	35.2	28.4	23.1	13.8
Reranker	33.8	27.5	22.4	13.2
Top-5	30.1	24.8	20.3	12.5
MMR	28.6	23.2	18.7	11.8
LLM-Sel.	25.9	20.8	16.5	10.4
Oracle	22.4	17.6	14.1	8.7

Table 12: Unsupported coverage rate (%): fraction of evidence-insufficient units ($\text{supp}(u_i, S_k) = 0$) that the response nevertheless addresses ($\text{cover}(u_i, R) = 1$). Higher values indicate greater reliance on parametric knowledge.

System	k	U-Rec	U-Prec	AC	Cite-Rec
DPR	3	40.1	38.6	58.2	80.5
	5	55.9	31.5	63.0	74.5
	10	70.8	22.3	65.8	66.4
Reranker	3	44.0	39.8	60.4	80.1
	5	59.8	33.2	65.2	78.4
	10	73.9	23.9	67.6	68.8
Top-5	3	48.2	47.1	62.8	84.7
	5	63.8	39.5	67.4	82.5
	10	78.3	28.4	69.5	76.8

Table 13: Effect of passage count k on TRIALIGN metrics with GPT-4o as generator. All values are percentages.

System	ROUGE-L	BERTSc.	AC
Random-5	28.4	62.1	47.5
BM25	31.7	64.8	52.2
DPR	35.2	67.3	63.0
Reranker	35.8	67.6	65.2
Top-5	37.9	69.1	67.4
MMR	37.1	68.4	69.1
LLM-Sel.	38.3	69.7	70.5
Oracle	39.1	69.2	71.8
Kendall τ with TRIALIGN *AC:*			
	0.93	0.86	–

Table 14: Aggregate text-overlap metrics vs. TRIALIGN AC, with GPT-4o as generator and $k = 5$. All values are percentages except Kendall τ.

System	GPT-4o	Qwen-72B	70B	8B
Random-5	47.3	43.1	40.1	30.7
BM25	51.7	45.5	41.5	29.0
DPR	62.4	53.9	48.5	29.5
Reranker	64.5	54.6	48.9	28.6
Top-5	66.7	57.3	50.9	29.1
MMR	68.1	57.7	50.8	27.0
LLM-Sel.	69.2	57.5	50.3	23.8
Oracle	69.4	55.4	48.7	21.3

Table 15: Verified-AC (%): fraction of RIUs that are both covered and correctly attributed, i.e., V-AC $= |\{u_i : \text{cover}(u_i, R){=}1 \wedge \text{attr}(u_i, R, S_k){=}1\}|/|U(Q)|$ ($k = 5$). Computed directly from unit-level coverage and attribution judgments; consistent with the failure attribution counts underlying Table 3.

relatively high AC even under weak evidence selection. Second, unsupported coverage decreases as selection quality improves (Random-5 to Oracle), because stronger selection leaves fewer evidence-insufficient units; among those that remain, support is absent from the selected evidence rather than from the full pool, often because the fixed k budget cannot include all supporting passages.

F.3 Context Size Ablation Across Selection Strategies

§4.5 reports the coverage–attribution trade-off for Top-5 (GTR) selection as k varies. Here we replicate the ablation across three selection strategies (DPR, Reranker, Top-5) to verify that the trade-off is not specific to GTR-style ranking (Table 13).

Across all three strategies, increasing k from 3 to 10 improves U-Rec by approximately 30 points and AC by approximately 7 points, while Cite-Rec decreases by 8–14 points. The coverage–attribution trade-off observed for Top-5 thus generalizes across rankers, confirming that it is not an artifact of a particular ranking method.

F.4 Quantitative Metric Comparison

Table 14 compares traditional metrics with TRIALIGN AC across the eight selection strategies using GPT-4o as generator. ROUGE-L and BERTScore preserve much of the same ranking signal as AC ($\tau \geq 0.86$), but neither metric can identify which stage is responsible for a failure. In contrast, TRIALIGN's failure attribution protocol distinguishes evidence insufficiency, answer under-coverage, and attribution misuse, providing diagnostic information that aggregate text-overlap metrics cannot.

F.5 Verified Answer Completeness

Answer completeness (AC) counts all covered units regardless of whether citations are correct. We define *Verified-AC* as the fraction of RIUs that are both covered *and* correctly attributed: V-AC $= |\{u_i : \text{cover}(u_i, R) = 1 \wedge \text{attr}(u_i, R, S_k) = 1\}|/|U(Q)|$. This metric penalizes systems that produce answers without proper citation support.

Tables 15 and 16 reveal a striking interaction

System	GPT-4o	Qwen-72B	70B	8B
Random-5	0.2	1.1	1.7	3.5
BM25	0.5	3.0	4.6	9.6
DPR	0.6	4.2	6.7	15.3
Reranker	0.7	4.9	7.6	17.2
Top-5	0.7	4.7	7.7	18.2
MMR	1.0	5.8	9.3	21.7
LLM-Sel.	1.3	7.3	11.2	25.8
Oracle	2.4	10.5	14.8	31.8

Table 16: Attribution loss: AC minus Verified-AC (percentage points). Represents the fraction of RIUs covered but lacking correct citations. The gap grows with both evidence quality and model weakness.

between evidence quality and citation capability. GPT-4o loses at most 2.4 points from attribution failures even under Oracle selection, whereas Llama-3.1-8B loses 31.8 points—more than half its nominal coverage lacks proper citation support. This gap motivates Verified-AC as a complementary metric for applications requiring verifiable answers (e.g., legal or medical QA), where citation correctness is as important as content coverage. The attribution loss increases monotonically with selection quality for all generators, but the effect is negligible for GPT-4o (at most 2.4 pp) and substantial for weaker models (up to 31.8 pp for Llama-8B), indicating that richer evidence contexts expose citation weaknesses that remain latent under sparse evidence.

F.6 Bootstrap Confidence Intervals

We report 95% bootstrap confidence intervals (10,000 resamples, percentile method) for the main metrics under Top-5 selection (Table 17). U-Rec and U-Prec intervals are computed from per-query evidence sufficiency scores; AC, Cite-Rec, and Cite-Prec intervals are computed from per-query judge scores.

The intervals confirm that most pairwise generator differences for AC and Cite-Rec are statistically significant (non-overlapping CIs). Qwen-72B and Llama-70B AC intervals overlap slightly, consistent with their similar parameter counts; all other pairwise comparisons are significant. All Cite-Rec differences are significant, consistent with citation capability being a stronger discriminator than content coverage (§4.2). Evidence-side metrics (U-Rec, U-Prec) have narrower CIs because they are computed directly from gold annotations without judge variance.

Table 18 provides CIs for the core quantitative claims made in §4.2–4.5, confirming their statistical robustness.

Metric	Generator	Mean [95% CI]
U-Rec@5	(all)	63.8 [61.4, 66.2]
U-Prec@5	(all)	39.5 [37.8, 41.4]
AC	GPT-4o	67.4 [65.4, 69.4]
	Qwen-72B	62.0 [59.8, 64.2]
	Llama-70B	58.6 [56.3, 60.9]
	Llama-8B	47.3 [44.7, 49.9]
Cite-Rec	GPT-4o	82.5 [80.2, 84.8]
	Qwen-72B	70.6 [68.1, 73.1]
	Llama-70B	63.7 [61.1, 66.3]
	Llama-8B	42.5 [39.5, 45.5]
Cite-Prec	GPT-4o	56.4 [53.4, 59.4]
	Qwen-72B	51.9 [48.6, 55.2]
	Llama-70B	49.1 [45.6, 52.6]
	Llama-8B	39.5 [35.6, 43.4]

Table 17: Bootstrap 95% confidence intervals for Top-5 (GTR) selection ($n = 887$ queries, 10,000 resamples). U-Rec/U-Prec are shared across generators (evidence-only). Non-overlapping CIs between generators confirm statistically significant differences.

Claim	95% CI
Context size ablation (Top-5, GPT-4o):	
AC gain $k = 3 \rightarrow 10$	+6.7 [+4.9, +8.5]
Cite-Rec drop $k = 3 \rightarrow 10$	−7.9 [−10.4, −5.4]
Failure share differences (§4.3):	
Evid%: Rand+4o − Top5+8B	+50.5 [+47.2, +53.8]
Evid%: BM25+4o − Orac+8B	+68.3 [+64.8, +71.8]
Verified-AC gap (Oracle, §4.3):	
V-AC: GPT-4o − 8B	+48.1 [+44.6, +51.6]
Attr. loss: 8B − GPT-4o	+29.4 [+26.8, +32.0]

Table 18: Bootstrap 95% CIs for key quantitative claims (10,000 resamples, percentile method, $n = 887$ queries). All differences are statistically significant at $\alpha = 0.05$.

F.7 Citation Generation Statistics

Table 19 reports the number of citation instances per response. Citation counts are extracted via regex matching of `[N]` markers.

Citation volume scales with both generator capability and selection quality. GPT-4o produces roughly twice as many citations as Llama-8B across all selection strategies. Interestingly, Oracle selection yields fewer citations than Top-5 despite higher evidence coverage. This suggests that when selected evidence is highly concentrated (few passages covering many RIUs), generators produce more concise, targeted citations rather than spreading attributions across many passages.

Selection	Generator	Mean	Median	Std
Top-5	GPT-4o	7.7	7	3.3
	Qwen-72B	6.5	6	2.9
	Llama-70B	5.9	5	2.7
	Llama-8B	4.0	4	1.8
Random-5	GPT-4o	5.1	5	2.4
	Qwen-72B	4.3	4	2.1
	Llama-70B	3.9	4	1.9
	Llama-8B	2.6	2	1.3
Oracle	GPT-4o	5.3	5	2.6
	Qwen-72B	4.5	4	2.2
	Llama-70B	4.1	4	2.0
	Llama-8B	2.8	3	1.4

Table 19: Citation instances per response (number of [N] markers). Stronger generators produce more citations and cite more diverse passages. Oracle has fewer citations than Top-5 because its selected passages are more concentrated on supporting evidence, reducing the need for citations to disambiguate.

G Diagnostic Case Study

We present a worked example that illustrates how the TRIALIGN diagnostic pipeline operates in practice. The query is taken in the spirit of ASQA's ambiguous factoid style: a short question with multiple legitimate interpretations whose complete answer must address each interpretation. To exhibit all three failure modes in a single example, the response below is constructed to be representative rather than drawn from any specific (selection, generator) configuration in our benchmark.

Query. *"Who has played James Bond in the official EON Productions films?"*

Required information units.

- u_1: Sean Connery as James Bond.
- u_2: George Lazenby as James Bond.
- u_3: Roger Moore as James Bond.
- u_4: Timothy Dalton as James Bond.
- u_5: Pierce Brosnan as James Bond.
- u_6: Daniel Craig as James Bond.

Evidence pool (simplified). From the 100-passage candidate pool, we show the support matrix for a selected set $S = \{e_2, e_3, e_4, e_5, e_9\}$. For clarity, citations in this example use passage subscript indices directly (i.e., [2] refers to e_2). Passages e_{11} and e_{15} (which would support u_5 and u_6 respectively) are present in the pool but not selected.

	u_1	u_2	u_3	u_4	u_5	u_6
e_2	1	0	0	0	0	0
e_3	0	0	1	1	0	0
e_4	0	1	0	1	0	0
e_5	1	0	0	0	0	0
e_9	0	0	0	0	0	0
e_{11} (not selected)	0	0	0	0	1	0
e_{15} (not selected)	0	0	0	0	0	1

Evidence sufficiency over S: supp(u_1, S)=1, supp(u_2, S)=1, supp(u_3, S)=1, supp(u_4, S)=1, supp(u_5, S)=0, supp(u_6, S)=0. Thus U-Rec@5 = 4/6 = 66.7% and U-Prec@5 = 4/5 = 80% (passage e_9 supports no RIU).

Generated response (illustrative, abbreviated).

> *Sean Connery originated the role of James Bond in the EON Productions films, starring in the first several entries of the series [2][5]. George Lazenby took over for a single film in the late 1960s [3]. Timothy Dalton later played the character in two films in the late 1980s [3][4].*

Diagnostic cascade for each unit.

- u_1 **(Connery):** supp=1, cover=1; cites [2][5], both of which support u_1 ⇒ attr=1. **Success.**
- u_2 **(Lazenby):** supp=1, cover=1; the response addresses u_2 but cites [3], which supports u_3 and u_4 rather than u_2. The correct citation would have been [4]. ⇒ attr=0. **Attribution Misuse.**
- u_3 **(Moore):** supp=1 via e_3, but the response does not mention Roger Moore at all ⇒ cover=0. Evidence was available but not used. **Answer Under-Coverage.**
- u_4 **(Dalton):** supp=1, cover=1; cites [3][4], both of which support u_4 ⇒ attr=1. **Success.**
- u_5 **(Brosnan):** supp=0 over S ⇒ **Evidence Insufficiency.** Passage e_{11} supports this unit but was not selected.
- u_6 **(Craig):** supp=0 over S ⇒ **Evidence Insufficiency.** Passage e_{15} supports this unit but was not selected. By the cascading protocol, both u_5 and u_6 are attributed to evidence selection regardless of whether the response mentions them from parametric knowledge, since grounded QA requires verifiable sources.

Diagnosis summary. This single example exhibits all three failure categories defined in §2.4: two evidence insufficiencies (u_5, u_6), one answer under-coverage (u_3), and one attribution misuse (u_2). Each diagnosis points to a different prescription:

- u_5, u_6: improve evidence selection so that passages e_{11} and e_{15} are included. A reranker or Oracle selector would surface these passages.
- u_3: improve generation coverage; supporting evidence was already available in S, so the bottleneck lies in the generator rather than evidence selection.
- u_2: improve claim-to-evidence mapping; the generator located the right information but attached it to the wrong citation index.

Aggregate metrics such as AC alone would report this response as "3 out of 6 covered" and conflate the three underlying causes. The cascading diagnostic protocol turns the same response into three distinct, actionable signals.