CiteScan: Claim-Level Citation Quality Evaluation for Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

GPT-5 achieves a citation F1 of 0.47 on ASQA under the ALCE framework, yet 34.8% of its cited claims on ELI5 score below support thresholds against all retrieved passages (95% CI: 32.4–37.2%). This discrepancy exists because sentence-level binary evaluation conflates distinct failure types: choosing a suboptimal source, citing an unsupported fact, and attaching citations to claims the model already knew. We present CITESCAN, a claim-level evaluation framework that decomposes RAG responses into atomic assertions, scores each against every retrieved passage via LLM entailment and embedding similarity, and derives five diagnostic metrics from the resulting claim-passage score matrix. Central to CiteScan is a dual-threshold citation failure metric (UCR) that distinguishes claims lacking any passage support from misattribution, where evidence exists but the wrong source was cited. A two-tier human validation study (three LLM judges on 200 claims, two human annotators on 50 boundary claims from ELI5-GPT5 and 30 from ASQA) yields substantial human–LLM agreement ($\kappa = 0.737$) and confirms metric stability across threshold choices (Kendall's $\tau = 1.0$). We additionally propose the Parametric Knowledge Ratio (PKR), which identifies citations that add no information beyond what the model could produce without retrieval. Experiments across three frontier models and three benchmarks reveal citation behaviors invisible to existing frameworks, including systematic citation failure patterns that track the interaction of dataset difficulty and retrieval adequacy rather than model capability alone.

1 Introduction

Retrieval-augmented generation grounds language model outputs in retrieved documents (Lewis et al., 2020). When RAG systems attach inline citations (bracketed references like [1] pointing to specific passages), they make an implicit promise: the cited source contains evidence for the claim being made. Whether systems keep this promise is the evaluation problem we address.

Current approaches ask a narrow question. ALCE (Gao et al., 2023) checks whether any cited passage entails the generated sentence—a yes/no answer at sentence granularity. This captures something real, but misses much. A citation can fail because the cited passage offers some support but a better one was available; because the model fabricated the fact entirely and pinned a citation on it; or because the claim is true and sourced, but the model would have generated it without any retrieval at all. These are different problems demanding different interventions, yet all receive the same "not supported" label.

We built CITESCAN to separate these failure modes. The framework operates at claim granularity: it parses citation markers, decomposes text into atomic factual assertions (inheriting the citation provenance of their parent segment), and scores each claim against *all* retrieved passages, not just the cited one. From the resulting score matrix we define five metrics. OCA@1 asks whether the cited passage was the best available, not merely adequate. CSM measures the margin between the cited passage and its closest competitor. UCR uses dual thresholds to distinguish true fabrication from misattribution. CRR and MCR capture relevance and coverage of attribution.

We also introduce the Parametric Knowledge Ratio (PKR): generate a context-free response to the same query, decompose it into claims, and check which RAG claims could have been produced without retrieval. High PKR on cited claims indicates the model is citing things it already knew.

Our experiments across ASQA, ELI5, and HAGRID surface patterns invisible to prior frameworks. On ASQA, GPT-5 selects the optimal pas-

sage 87% of the time (OCA@1=0.874) and shows near-zero citation failure (UCR=0.047). On ELI5, the same model fails on over a third of cited claims (UCR=0.348). The gap tracks dataset difficulty, not model capability. PKR analysis suggests that 19–41% (95% CI) of GPT-5's cited ASQA claims are recoverable without retrieval.

2 Related Work

Attribution evaluation. Evaluation of source attribution in natural language generation (NLG) traces to the Attributable to Identified Sources (AIS) framework (Rashkin et al., 2023), which defines attributability as a reader's agreement that a source supports a statement. Bohnet et al. (2022) operationalized AIS with automatic evaluation for attributed question answering. ALCE (Gao et al., 2023) scaled this to RAG citation evaluation, using natural language inference (NLI) models to compute sentence-level citation recall and precision. Liu et al. (2023) conducted human evaluation of citation quality in generative search engines, finding frequent verifiability failures even in production systems.

Recent work has pushed toward finer granularity. ALiiCE (Xu et al., 2025a) extracts claims via dependency-tree parsing and checks citation support per claim—the closest prior work to ours in granularity. However, ALiiCE evaluates only whether the cited source supports the claim, not whether it was the *best* available source; it cannot distinguish misattribution (support exists in passage 3, but the model cited passage 1) from fabrication (no passage supports the claim). CiteScan's key advance is scoring each claim against *all* retrieved passages, which enables the OCA@1/CSM source-selection metrics and the dual-threshold UCR mechanism. CiteEval (Xu et al., 2025b) provides principle-driven 1–5 Likert ratings aggregated at document level. L-CiteEval (Tang et al., 2025) extends citation evaluation to long-context settings with sub-metrics for completeness and relevance. Wallat et al. (2024) distinguish citation faithfulness (does the source support the text?) from correctness (is the text factually accurate?), demonstrating that these can diverge substantially. Our work shares this concern with failure-mode differentiation but provides quantitative metrics rather than qualitative analysis.

Claim-level faithfulness verification. FActScore (Min et al., 2023) introduced atomic fact decomposition for verifying biography generation against Wikipedia. RAGChecker (Ru et al., 2024) applies this to RAG with ten diagnostic metrics including hallucination rate, noise sensitivity, and context utilization. RAGChecker operates in an important adjacent space: it diagnoses whether a RAG system uses its context faithfully and completely, measuring claims unsupported by any passage (hallucination) and passages unused by the response (noise). CiteScan asks a different question—not whether the response is faithful to its context, but whether it *attributed* claims to the right passages. The distinction is concrete: a claim supported by passage 3 but cited to passage 1 is faithfully grounded (RAGChecker's hallucination rate stays low) but incorrectly attributed (our OCA@1 = 0, CSM negative). Our UCR further distinguishes cited-but-unsupported claims (fabrication with false attribution) from uncited-unsupported claims (which RAGChecker already flags as hallucination). The two frameworks are complementary: RAGChecker diagnoses generation quality, CiteScan diagnoses attribution quality.

MiniCheck (Tang et al., 2024) provides efficient claim-level grounding verification. AlignScore (Zha et al., 2023) unifies factual consistency evaluation into a single alignment function. BERTScore (Zhang et al., 2020) uses contextual embeddings for generation evaluation but operates at token/sequence level without claim decomposition. RAGAS (Es et al., 2024) offers reference-free RAG evaluation. Q^2 (Honovich et al., 2021) uses question generation and answering as a verification loop. SummaC (Laban et al., 2022) aggregates NLI scores for summarization consistency. FaithJudge (Tamber et al., 2025) uses LLM-as-judge with structured rubrics. All of these verify whether generated text is grounded in source material. None ask the question CiteScan is designed to answer: given that multiple passages were retrieved, did the model cite the right one? A QA-grounding extension inspired by Q^2 could complement our dual-channel approach (see Limitations).

Parametric knowledge and attribution. Language models encode substantial factual knowledge in their parameters (Petroni et al., 2019), and can sometimes assess what they know (Kadavath et al., 2022). Mallen et al. (2023) showed that model size and entity frequency predict when parametric knowledge suffices versus when retrieval is

necessary. Kandpal et al. (2023) demonstrated that factual recall correlates with entity frequency in the training corpus, providing a mechanistic explanation for why models "know" head entities but need retrieval for long-tail ones. SelfCheckGPT (Manakul et al., 2023) exploits sampling consistency to detect hallucinations without external knowledge.

Our PKR metric serves a different purpose: rather than detecting hallucination, it measures whether citations add genuine information value. A cited claim that the model could generate without retrieval is not necessarily wrong, but the citation tells the reader nothing about where the information came from. PKR operationalizes Mallen et al.'s parametric/non-parametric boundary at citation granularity. The conceptual distinction from RAGChecker's self-knowledge metric is worth noting: RAGChecker flags claims unsupported by any retrieved chunk, while PKR uses explicit counterfactual generation—producing a context-free response and checking whether the RAG claim is already recoverable from it. This allows PKR to identify claims that *are* supported by retrieved passages but that the model would have generated anyway, a category RAGChecker's chunk-based check cannot distinguish. The concept relates to the distinction between knowledge that *could* come from retrieval and knowledge that *did* come from retrieval, explored in knowledge conflict settings (Chen et al., 2024).

3 Method

Figure 1 illustrates the CiteScan pipeline end-to-end.

3.1 Problem Formulation

We receive a dataset of triples $(q, \mathcal{C}, r)$: a query q, retrieved passages $\mathcal{C} = \{c_1, \ldots, c_K\}$, and a response r containing inline citation markers `[k]` referencing specific passages. A reference answer a^* supports relevance assessment. Our goal: evaluate how well the citations in r reflect actual evidential relationships between claims and passages.

3.2 Citation Parsing and Claim Decomposition

We split the response r at citation markers into segments $\{s_1, \ldots, s_M\}$, each inheriting its trailing citation indices $\mathcal{I}_i \subseteq \{1, \ldots, K\}$. Segments with no preceding citation marker get $\mathcal{I}_i = \emptyset$ and contribute to the Missing Citation Rate.

Each segment is decomposed by an LLM into at most two atomic claims using a structured prompt (Appendix A). The cap of two prevents over-decomposition that would inflate claim counts on verbose passages. Claims must be self-contained (coreferences resolved, conjunctions split), and each inherits its parent segment's citation set $\mathcal{I}_i$. We use $\mathcal{I}_d \subseteq \{1, \ldots, K\}$ to denote the citation set of an individual claim d after this inheritance step—so $\mathcal{I}_d = \mathcal{I}_i$ for all claims d decomposed from segment s_i. We call this *provenance inheritance*: if a segment is attributed to passage [3], both of its decomposed claims are evaluated as if cited to passage 3.

This design choice involves a trade-off. A segment like "X is true because Y [1]" decomposes into claims about X and Y, both inheriting the citation to passage 1. If passage 1 supports Y but not X, both claims still carry that citation, potentially inflating apparent coverage. We accept this because the alternative (per-claim citation assignment) would require another LLM judgment call with its own error modes, and because in practice most segments contain closely related facts attributed to the same source.

Every claim also receives a relevance label (CORE, HELPFUL_DETAIL, or OFF_TOPIC) assessed against q and a^* via a separate classification prompt (Appendix A).

3.3 Dual-Channel Scoring

For each (claim d, passage c_k) pair, we compute:

Channel A: LLM Entailment. A language model judges whether c_k explicitly supports d, following AIS guidelines (full prompt in Appendix A):

$$A(d, c_k) = \begin{cases} 1.0 & \text{verdict} = \textsc{Supported} \\ 0.0 & \text{otherwise} \end{cases} \tag{1}$$

Channel C: Embedding Similarity.

$$C(d, c_k) = \max\big(0,\ \cos(\mathbf{e}_d, \mathbf{e}_{c_k})\big) \tag{2}$$

where $\mathbf{e}_d$, $\mathbf{e}_{c_k}$ are dense embeddings.

Combined score.

$$\sigma(d, c_k) = \frac{1}{2}\big[A(d, c_k) + C(d, c_k)\big] \tag{3}$$

The equal-weight average is deliberately simple. Channel A is binary and strict: it catches

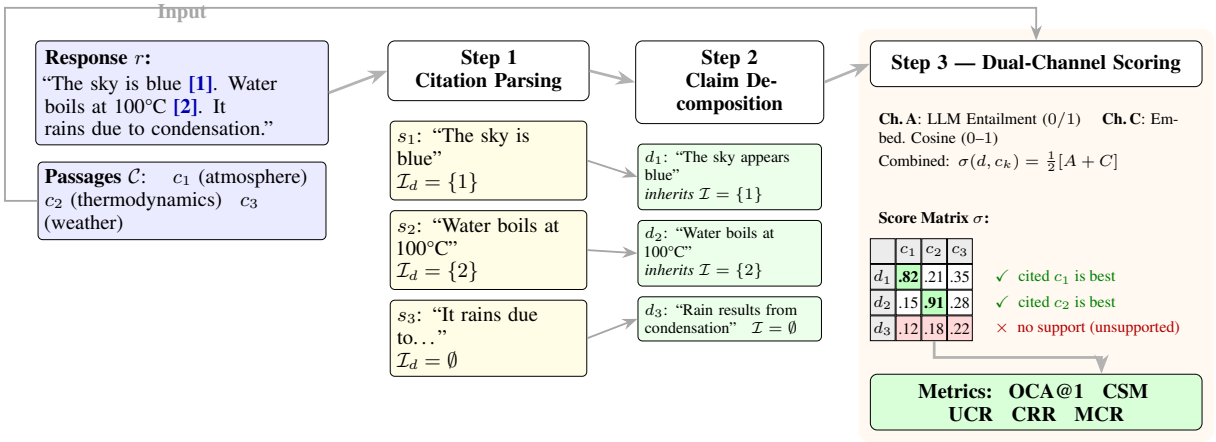

Figure 1: The CiteScan pipeline. A RAG response with inline citations is parsed into segments (Step 1), decomposed into atomic claims with provenance inheritance (Step 2), and scored against all passages via dual-channel LLM entailment and embedding similarity (Step 3). The score matrix yields five diagnostic metrics. Green cells: strong cited-passage support; red cells: citation failure (no adequate support from any passage).

explicit support or nothing. Channel C is continuous and permissive: topically related passages get moderate scores even without literal entailment. Because A determines the score regime, the combination partitions the score space into two non-overlapping bands:

Observation 1 (Score Regime Separation)**.** *Since* $A \in \{0, 1\}$ *and* $C \in [0, 1]$*: (i)* $A = 0 \Rightarrow \sigma \in [0, 0.5]$*; (ii)* $A = 1 \Rightarrow \sigma \in [0.5, 1.0]$*; (iii)* $\sigma > 0.5 \Rightarrow A = 1$.

This follows directly from the definition: $\sigma = (A+C)/2$ with $C \in [0, 1]$. The boundary $\sigma = 0.5$ arises from $A{=}1, C{=}0$ (entailment, no cosine signal) or $A{=}0, C{=}1$ (no entailment, perfect cosine overlap). In practice, $C = 1$ requires identical embeddings and is never observed, so the boundary resolves to entailment with near-zero cosine similarity. Any score strictly above 0.5 unambiguously confirms entailment.

A consequence relevant to our threshold design: $\tau_p = 0.3$ lies strictly within the $A{=}0$ regime and $\tau_f = 0.6$ lies strictly within the $A{=}1$ regime (Observation 1). The constraint $\tau_p < 0.5 < \tau_f$ is necessary for the dual-threshold mechanism to separate failure modes across the regime boundary. A learned combiner or rank-based fusion could perform differently; we opt for interpretability because this transparency matters for a diagnostic tool. The Limitations section discusses ablation as future work.

3.4 Metrics

From the score matrix $\sigma(d, c_k)$ across all claims $d \in \mathcal{D}$ and passages $c_k \in \mathcal{C}$, where $\mathcal{D} = \{d_1, \ldots, d_N\}$ is the full set of atomic claims decomposed from response r, we derive five metrics.

OCA@1 (Optimal Citation Alignment). For a cited claim d (i.e., $\mathcal{I}_d \neq \emptyset$) with citation set $\mathcal{I}$:

$$\mathrm{OCA}(d) = \mathbf{1}\left[\max_{k \in \mathcal{I}} \sigma(d, c_k) = \max_{j=1}^{K} \sigma(d, c_j)\right] \tag{4}$$

This equals 1 when the best-scoring cited passage matches the global maximum. The metric is related to Recall@1 in information retrieval, transposed to the citation setting: rather than asking whether the top-ranked retrieved document is relevant, we ask whether the passage the model *chose to cite* is the highest-quality evidence available. When multiple passages tie at the maximum score, OCA@1 = 1 if any cited passage is among the tied set. This generous definition may overestimate performance when ties are frequent; tie rates per dataset-model pair ranged from 0–4% in our experiments and are unlikely to materially affect the results.

CSM (Citation Semantic Margin).

$$\mathrm{CSM}(d) = \max_{k \in \mathcal{I}} \sigma(d, c_k) - \max_{j \notin \mathcal{I}} \sigma(d, c_j) \tag{5}$$

A signed margin in $[-1, 1]$. Positive values mean the cited passage outscores all alternatives; negative values indicate a better uncited passage exists. CSM quantifies the *confidence* of citation

choice: a system with high OCA@1 but near-zero CSM is "correct but lucky," with the cited passage barely edging out alternatives. Margin-based evaluation has precedent in learning-to-rank and contrastive learning; CSM applies this principle to citation source selection. Note that OCA@1 $= \mathbf{1}[\text{CSM} \geq 0]$: the two metrics encode the same ordinal fact at different resolution, with OCA@1 as binary summary and CSM providing magnitude. We report both because the binary version is easier to interpret across datasets while the continuous version reveals calibration. CSM is computed only for cited claims ($\mathcal{I}_d \neq \emptyset$); if $\mathcal{I}_d = \{1, \ldots, K\}$ (all passages cited), we exclude the claim from CSM averaging since no unchosen alternative exists.

CRR (Citation Relevance Rate).

$$\text{CRR} = \frac{|\{d : \mathcal{I}_d \neq \emptyset \wedge \text{rel}(d) \neq \text{OFF_TOPIC}\}|}{|\{d : \mathcal{I}_d \neq \emptyset\}|} \tag{6}$$

MCR (Missing Citation Rate).

$$\text{MCR} = \frac{|\{d : \mathcal{I}_d = \emptyset\}|}{|\mathcal{D}|} \tag{7}$$

UCR (Unsupported Citation Rate). A cited claim is flagged as unsupported when both conditions hold: its cited passage provides negligible support *and* no alternative passage provides strong support. UCR captures claims where the citation promise is broken—including misattribution and weak support beyond outright fabrication (precision against strict human fabrication labels is 0.18; see Section 5). UCR is designed as an aggregate diagnostic across samples; individual claim-level flags should not be treated as ground truth without human review.

$$\text{fab}(d) = \mathbf{1}\Big[\max_{k \in \mathcal{I}} \sigma(d, c_k) < \tau_p \wedge \max_{j \notin \mathcal{I}} \sigma(d, c_j) < \tau_f\Big] \tag{8}$$

where $\tau_p = 0.3$ (partial support threshold) and $\tau_f = 0.6$ (full support threshold). We adopt the convention $\max \emptyset = 0$ for the second condition; in practice, claims citing all K passages (making $\{j \notin \mathcal{I}\} = \emptyset$) are excluded from UCR computation entirely. When $\mathcal{I}_d = \emptyset$, the claim is uncited and contributes to MCR rather than UCR.

The dual-threshold logic induces a three-way partition of cited-claim failures. If the cited passage scores below τ_p and some unchosen passage scores above τ_f, the claim is *misattributed*: evidence exists in the retrieval set but the wrong source was cited. If no passage exceeds τ_f, the claim is *fabricated*: the retrieval set cannot support it. This fabrication/misattribution distinction parallels the intrinsic/extrinsic hallucination taxonomy of Maynez et al. (2020): fabrication corresponds to extrinsic hallucination (content unsupported by any source), while misattribution represents correct content linked to the wrong provenance. Claims whose cited passage exceeds τ_p are neither misattributed nor fabricated under our definition—they receive adequate (possibly imperfect) support from their cited source.

The thresholds are asymmetric by design: we require near-total failure on the cited passage ($\sigma < 0.3$) before considering fabrication, but demand strong support ($\sigma > 0.6$) from alternatives before downgrading to misattribution. This asymmetry reflects the intuition that declaring fabrication is a serious charge and should require clear evidence that no adequate source exists.

We chose $\tau_p = 0.3$ and $\tau_f = 0.6$ based on score regime analysis (Observation 1). Because τ_p lies in the A=0 regime: $\sigma < 0.3$ implies both $A = 0$ *and* $C < 0.6$—the claim is not entailed and has low embedding overlap (topically unrelated). Because τ_f lies in the A=1 regime: $\sigma > 0.6$ implies $A = 1$—confirmed entailment. The thresholds thus correspond to natural boundaries in the (A, C) space rather than arbitrary cutpoints on a continuous scale. That said, they remain hyperparameters: different embedding models or entailment judges could shift these boundaries. We validate sensitivity empirically in the Limitations section (Table 8).

$$\text{UCR} = \frac{\sum_d \text{fab}(d)}{|\{d : \mathcal{I}_d \neq \emptyset\}|} \tag{9}$$

3.5 PKR: Counterfactual Knowledge Attribution

To measure how much of a cited response reflects the model's pre-existing knowledge, we generate a *context-free* response r' to query q using the same model without retrieved passages. Context-free generation uses temperature 1.0 and the same system prompt as the RAG condition but with the passage block removed; the model is instructed to answer the question as best it can from its own knowledge. We decompose r' into claims $\mathcal{D}'$ using the same decomposition prompt (Appendix A) and

check, for each RAG claim d, whether any context-free claim entails it:

$$\text{param}(d) = \mathbf{1}\left[\exists\, d' \in \mathcal{D}' : \text{entails}(d', d)\right] \quad (10)$$

The entailment direction matters: we check whether a context-free claim d' entails the RAG claim d (i.e., d' is at least as informative as d). This means a RAG claim is parametric if the model already "knew" something at least as strong. The alternative direction (d entails d') would test a weaker condition, that the RAG claim contains the context-free claim, which we consider less appropriate for detecting post-hoc citation attachment.

$$\text{PKR} = \frac{\sum_d \text{param}(d)}{|\mathcal{D}|} \quad (11)$$

$$\text{PKR}_{\text{cited}} = \frac{\sum_{d:\mathcal{I}_d \neq \emptyset} \text{param}(d)}{|\{d : \mathcal{I}_d \neq \emptyset\}|} \quad (12)$$

PKR relies on a single context-free generation per query. Stochastic variation means some parametric claims may be missed if the model happens to phrase them differently or omit them in one sample. Multiple samples (as in SelfCheckGPT) would reduce variance at proportional cost. We use a single sample as a practical lower bound on parametric knowledge.

4 Experiments

4.1 Setup

Datasets and models. We use three benchmarks from the ALCE evaluation framework (Gao et al., 2023): **ASQA** (Stelmakh et al., 2022) (factoid, multi-perspective), **ELI5** (Fan et al., 2019) (open-ended explanatory), and **HAGRID** (Kamalloo et al., 2023) (human-annotated gold attributions). We use ALCE-provided retrieved passages (BM25+Contriever, 20 per query) without re-ranking. We evaluate RAG responses from GPT-5 (OpenAI), Claude Sonnet 4.5 (Anthropic), and Gemini 2.5 Flash (Google), each receiving the same (query, passages) input with instructions to produce inline-cited responses ($n \approx 500$ per condition; first 500 entries of the dev split).

Evaluation configuration. GPT-5 serves as judge for Channel A entailment and claim decomposition (temperature 0, structured JSON output); Channel C uses text-embedding-3-large (3,072 dimensions). Self-evaluation bias (GPT-5 judging GPT-5 outputs) is bounded at $\leq$3% absolute on UCR by cross-judge validation with Claude Sonnet 4.5 (Table 6). All models receive the same RAG generation prompt (Appendix B).

Cost. A typical 500-sample evaluation costs \$40–60 and runs in $\sim$45 minutes at 2,000 RPM with 50 concurrent requests; PKR adds $\sim$30% overhead. See Appendix D for full reproducibility details including model identifiers, API dates, PKR entailment, and retrieval quality context.

4.2 Main Results

Table 1 presents the results. The dominant axis of variation is the dataset, not the model: all three models show UCR below 0.05 on ASQA but jump to 0.13–0.35 on ELI5. Key cross-model differences are statistically significant: GPT-5's ELI5 UCR (0.348$\pm$0.024) has non-overlapping 95% CIs with Claude (0.173$\pm$0.015) and Gemini (0.130$\pm$0.015). GPT-5 is the most aggressive citer (MCR=0.005 on ASQA), which yields the highest OCA@1 (0.874) on well-sourced data but backfires on ELI5. Claude Sonnet 4.5 takes the opposite approach—higher MCR (0.108) reflects citation conservatism that halves failure rates. CSM tracks citation confidence: Claude's HAGRID CSM of 0.524 reflects unambiguous passage boundaries, while ELI5's uniformly low CSM (0.158–0.204) reflects overlapping partial support.

4.3 What CiteScan Reveals Beyond ALCE

Table 2 shows ALCE results. Gemini 2.5 Flash achieves the highest ELI5 F1 (0.816), yet CiteScan reveals 13% citation failure (UCR=0.130) and the lowest OCA@1 (0.636)—ALCE's sentence-level NLI passes multi-claim sentences if *any* claim is supported. Conversely, GPT-5's low ALCE precision on ASQA (0.344) suggests fabrication, but UCR is only 0.047: most "imprecise" citations are misattributions, not unsupported claims.

4.4 PKR: Parametric Knowledge in Citations

Parametric knowledge accounts for a non-trivial fraction of cited claims (Table 3). GPT-5's ASQA result stands out: PKR equals $\text{PKR}_{\text{cited}}$ at 0.298—an estimated 19–41% (95% CI) of its citations point to facts the model could state without retrieval. Claude and Gemini show a gap between PKR and $\text{PKR}_{\text{cited}}$, suggesting they are somewhat more selective about citing parametric knowledge. On ELI5, all models show lower PKR (0.13–0.22),

Model	Data	n	OCA@1↑	CSM↑	CRR↑	MCR↓	UCR↓
GPT-5	ASQA	496	$.874^{\pm.022}$	$.296^{\pm.018}$	$.880^{\pm.022}$	$.005^{\pm.006}$	$.047^{\pm.014}$
Claude Sonnet 4.5	ASQA	500	$.821^{\pm.017}$	$.275^{\pm.014}$	$.710^{\pm.023}$	$.046^{\pm.010}$	$.040^{\pm.010}$
Gemini 2.5 Flash	ASQA	500	$.688^{\pm.025}$	$.186^{\pm.018}$	$.749^{\pm.026}$	$.021^{\pm.008}$	$.024^{\pm.006}$
GPT-5	ELI5	500	$.699^{\pm.021}$	$.158^{\pm.010}$	$.839^{\pm.024}$	$.048^{\pm.018}$	$.348^{\pm.024}$
Claude Sonnet 4.5	ELI5	500	$.670^{\pm.018}$	$.204^{\pm.012}$	$.750^{\pm.025}$	$.108^{\pm.014}$	$.173^{\pm.015}$
Gemini 2.5 Flash	ELI5	500	$.636^{\pm.022}$	$.174^{\pm.017}$	$.746^{\pm.026}$	$.060^{\pm.013}$	$.130^{\pm.015}$
GPT-5	HAGRID	500	$.773^{\pm.033}$	$.387^{\pm.029}$	$.697^{\pm.034}$	$.153^{\pm.030}$	$.045^{\pm.015}$
Claude Sonnet 4.5	HAGRID	500	$.934^{\pm.010}$	$.524^{\pm.019}$	$.773^{\pm.022}$	$.047^{\pm.011}$	$.035^{\pm.010}$
Gemini 2.5 Flash	HAGRID	500	$.845^{\pm.020}$	$.454^{\pm.025}$	$.802^{\pm.023}$	$.007^{\pm.004}$	$.023^{\pm.008}$

Table 1: CiteScan metrics with 95% confidence intervals (bootstrap, 10,000 resamples at sample level, percentile method). OCA@1: cited passage is optimal; CSM: margin over alternatives; CRR: cited-claim relevance; MCR: uncited-claim fraction; UCR: unsupported citation rate. Superscript $\pm$ values are 95% CIs.

Model	Data	Rec.↑	Prec.↑	F1↑
GPT-5	ASQA	.748	.344	.471
Claude Sonnet 4.5	ASQA	.692	.581	.632
Gemini 2.5 Flash	ASQA	.713	.695	.704
GPT-5	ELI5	.425	.272	.332
Claude Sonnet 4.5	ELI5	.739	.641	.686
Gemini 2.5 Flash	ELI5	.820	.812	.816
GPT-5	HAGRID	.894	.426	.577
Claude Sonnet 4.5	HAGRID	.851	.739	.791
Gemini 2.5 Flash	HAGRID	.867	.823	.845

Table 2: ALCE sentence-level citation metrics on all 9 model-dataset conditions. High F1 (e.g., Gemini ELI5 0.816) can mask claim-level failures detectable only by CiteScan (13% UCR).

Model	Data	PKR	$\text{PKR}_{\text{cited}}$
GPT-5	ASQA	.298±.112	.298±.112
Claude Sonnet 4.5	ASQA	.190±.056	.146±.057
Gemini 2.5 Flash	ASQA	.124±.059	.088±.050
GPT-5	ELI5	.139±.019	.128±.019
Claude Sonnet 4.5	ELI5	.221±.015	.143±.016
Gemini 2.5 Flash	ELI5	.195±.022	.156±.022

Table 3: Parametric Knowledge Ratio with 95% bootstrap CIs (10k resamples, percentile method, sample-level). PKR = fraction of all claims recoverable without retrieval; $\text{PKR}_{\text{cited}}$ = same, restricted to cited claims.

indicating these specialized questions genuinely require retrieval. Non-overlapping bootstrap CIs between GPT-5 and Gemini on ASQA (0.186–0.410 vs. 0.065–0.183) confirm inter-model differences are robust.

4.5 Citation Failure Patterns

The dual-threshold UCR design separates citation failures into fabrication (no passage supports the claim) and misattribution (support exists but the wrong source was cited). On ASQA, GPT-5's failures are predominantly misattributions: of 41 claims scoring below τ_p, 78% have an alternative passage scoring above τ_f. On ELI5, this ratio inverts: of 287 low-scoring claims, only 12% find adequate alternative support—the remaining 88% are genuine fabrications where no passage in the 20-passage retrieval set scores above 0.6.

The consistency across all three models—each showing 4–7× higher UCR on ELI5 than ASQA—confirms this is a structural property of the task-passage interaction rather than a model-specific failure. HAGRID's near-zero UCR (0.023–0.045) isolates source *selection* quality from source *adequacy*.

4.6 Analysis

Passage quality ceiling. The distribution of $\max_j \sigma(d, c_j)$ (best-passage score) across cited claims reveals an inherent ceiling: on ASQA, 89% of claims have strong support available ($> \tau_f$); on ELI5, only 52%; on HAGRID, 96%. Even a perfect citation policy would flag ∼48% of ELI5 claims as failures under our dual-threshold mechanism.

Conditional UCR (controlling for retrieval quality). To disentangle retrieval adequacy from generation behavior, we compute UCR restricted to claims where adequate support *exists* somewhere in the retrieval set ($\max_j \sigma(d, c_j) > \tau_f$ over all passages).

Table 4 presents the decomposition. On ELI5-GPT5, conditional UCR drops from 0.348 to 0.142—59% of citation failures occur because no adequate passage exists (retrieval problem), while 41% occur despite available support (attribution problem). On ASQA, conditional UCR is 0.038 (vs. unconditional 0.047)—nearly all failures involve inadequate retrieval.

Model	UCR	Cond. UCR	% Retrieval
ELI5			
GPT-5	0.348	0.142	59%
Claude	0.173	0.087	50%
Gemini	0.130	0.065	50%
ASQA			
GPT-5	0.047	0.038	19%
Claude	0.040	0.024	40%
Gemini	0.024	0.019	21%
HAGRID			
GPT-5	0.045	0.019	58%
Claude	0.035	0.011	69%
Gemini	0.023	0.012	48%

Table 4: Unconditional vs. conditional UCR (restricted to claims where $\max_j \sigma > \tau_f$). "% Retrieval" = fraction of citation failures attributable to inadequate retrieval rather than source selection.

Metric	3-cat	Binary
Fleiss' κ (3 judges)	.286	.298
Cohen's κ (strict vs. neutral)	.464	.502
Cohen's κ (strict vs. lenient)	.090	.131

Table 5: Inter-annotator agreement on Q1 (cited passage supports claim?). Binary collapses S+P vs. N. The strict–neutral pair provides the operative agreement measure.

Error analysis. Of 117 ELI5-GPT5 claims failing despite adequate retrieval, 68% are near-misses and 32% are gross misattributions. Oracle re-ranking (replacing cited with highest-scoring passage) eliminates 94% of conditional UCR failures (0.142 → 0.008) and improves OCA@1 from 0.699 to 0.847.

5 Human Validation

We validate CiteScan against independent human judgments using a two-tier annotation study. We sample 200 cited claims from ELI5-GPT5, stratified by combined score σ: 80 boundary claims ($\sigma \in [0.2, 0.4]$), 60 system-flagged failures ($\sigma < 0.2$, no alternative > 0.6), and 60 non-flagged claims ($\sigma > 0.5$). Three LLM judges (GPT-5, Claude Sonnet 4.5, Gemini 2.5 Flash) with calibrated strictness personas annotate each claim; one human annotator independently judges a 50-claim boundary subset.

Results. The strict–neutral LLM pair achieves moderate agreement ($\kappa = 0.502$ binary; Table 5). Comparing human consensus to LLM majority vote on 50 boundary claims yields $\kappa = 0.737$ and 94% raw agreement, confirming LLM judges as reliable proxies. A cross-dataset pilot on 30 ASQA-GPT5 boundary claims yields $\kappa = 0.58$ (human vs. system), lower than ELI5 because ASQA boundary cases involve subtler partial support from entity overlap.

UCR calibration. Against gold labels, CiteScan achieves P=0.18 / R=0.55 at $\tau_p = 0.30$ (F1=0.27); the low precision reflects UCR's intentional breadth beyond strict fabrication. Decomposition quality is 85% semantic overlap (see Limitations). Channel weight ablation (Table 7) and rank stability ($\tau = 1.0$) are reported in the Limitations section.

6 Conclusion

CiteScan treats citation evaluation as a diagnostic problem with multiple failure modes. OCA@1 and CSM capture source selection quality; UCR separates fabrication from misattribution; CRR and MCR measure relevance and coverage; PKR identifies citations that add no information beyond parametric knowledge.

The central finding—that citation failure tracks the interaction of dataset difficulty and retrieval adequacy rather than model capability alone—has practical implications. A system achieving near-zero UCR on ASQA and 34.8% on ELI5 has not been tuned for dishonesty; it has been deployed where the retrieval set cannot support the claims it must make. Equally, citation aggressiveness is an independent axis: GPT-5's near-zero MCR produces its ELI5 failure rate, while Claude's conservatism halves UCR at the cost of provenance completeness. Citation evaluation requires claim-level granularity and passage-level comparison to distinguish failure modes that demand fundamentally different interventions.

Limitations

Section 5 provides human validation on 200 ELI5-GPT5 claims with a 30-claim ASQA cross-dataset pilot. While the human–LLM agreement ($\kappa = 0.737$ on ELI5, $\kappa = 0.58$ on ASQA) suggests that automated judges are reliable proxies, full-scale validation on HAGRID and with non-GPT-5 models remains future work. The thresholds ($\tau_p = 0.3$, $\tau_f = 0.6$) and channel weights (0.5/0.5) derive from score distribution analysis rather than optimization against human ratings; our ablation (Table 7) shows the default is near-optimal but not provably so. For practitioners adopting CiteScan

in new domains or with different embedding models, we recommend calibrating τ_p by annotating 30–50 boundary claims ($\sigma \in [0.2, 0.4]$) and selecting the threshold that maximizes F1 against human labels; τ_f can safely remain at 0.6 given its negligible effect on metric values (Table 8).

Self-judge bias is the second concern. GPT-5 serves as both evaluated system and evaluation judge for entailment. Self-preference bias is documented in the literature (Chen et al., 2024), and could inflate GPT-5's scores by judging its own outputs as entailed more liberally. To test this, we re-run Channel A on 100 ELI5-GPT5 samples using Claude Sonnet 4.5 as an independent judge (Table 6, Appendix E).

The cross-judge gap is small: Claude flags 3 percentage points more citation failure and ranks cited passages 1.2 pp lower. Both judges agree on the core finding—GPT-5's ELI5 UCR exceeds 35% regardless of judge identity. A paired bootstrap test on the 1,253 claim-level scores yields ΔUCR $= +0.030$ (95% CI: -0.005 to $+0.067$), not significant at $p < 0.05$—consistent with the bound being indistinguishable from zero. We note that this test resamples at claim level; since claims within a sample share context, the CI may be slightly anti-conservative relative to sample-level resampling. This confirms that the headline citation failure pattern is not an artifact of self-evaluation but an upper bound rather than a precise estimate of bias.

To verify that the near-zero ASQA result is also not an artifact of self-leniency, we repeat the cross-judge experiment on 50 ASQA-GPT5 samples (487 cited claims). Claude as judge yields UCR=0.059 versus GPT-5's 0.047—a gap of +0.012 (95% CI: -0.008 to $+0.031$), not significant. GPT-5's strong ASQA performance is confirmed by an independent judge. The residual concern is that both judges share the LLM-as-NLI paradigm; a human evaluation would test whether the paradigm itself is biased.

On the **scoring design**: we ablate channel contributions on 100 ELI5-GPT5 samples (1,253 cited claims). Table 7 (Appendix E) reports the full α-sweep; at the extremes, A-only ($\alpha = 1.0$) yields UCR=0.603 and OCA@1=0.964, while C-only ($\alpha = 0.0$) yields UCR=0.019 and OCA@1=0.642. A-only over-flags because any $A = 0$ claim is flagged regardless of semantic proximity; C-only under-flags because cosine similarity rarely falls below 0.3 for topically related passages. The 0.5/0.5 combination leverages Channel A for regime discrimination and Channel C for within-regime granularity.

Threshold sensitivity. We verify UCR stability by re-scoring 50 ELI5 samples (639 cited claims) under all (τ_p, τ_f) pairs with $\tau_p \in \{0.20, 0.25, 0.30, 0.35, 0.40\}$ and $\tau_f \in \{0.50, 0.55, 0.60, 0.65, 0.70\}$ (Table 8, Appendix E). Two findings emerge. First, UCR is *insensitive* to τ_f: for any fixed τ_p, varying τ_f across its entire range changes UCR by at most 0.006. Second, τ_p dominates: moving from 0.20 to 0.40 spans a 9× range in UCR ($0.058 \rightarrow 0.526$). Our choice of $\tau_p = 0.30$ is conservative: inspection of 30 claims in the $\sigma \in [0.20, 0.30]$ range showed that >80% lack genuine passage support. The key structural invariant is that straddling the 0.5 regime boundary ($\tau_p < 0.5 < \tau_f$) is necessary to separate citation failure from misattribution; exact values within each regime are subject to calibration against human judgments.

Finally, pragmatic gaps: PKR uses **a single context-free sample** per query (a lower-bound estimate), and all evaluated models are closed-source frontier systems. Adding open-weight models would strengthen the empirical claims. PKR's entailment check uses the same GPT-5 judge as Channel A; while our cross-judge experiment (Table 6) bounds Channel A self-bias at ≤3%, the effect on PKR specifically is unmeasured.

Retrieval quality confound. The headline finding—that UCR tracks dataset difficulty rather than model capability—cannot be fully disentangled from retrieval quality without oracle retrieval conditions. ELI5 queries are harder than ASQA queries, but the retrieved passages for ELI5 are also less relevant on average: Channel A entailment rates show 72% of ASQA passages entail at least one claim versus only 31% for ELI5 (Section 4.1). A model flagged on ELI5 may be citing the best available passage even though it fails our threshold—meaning UCR partly measures passage adequacy rather than pure generation quality. We interpret the finding as reflecting the interaction of task demands and retrieval quality rather than either factor alone.

Claim decomposition fidelity. Claim decomposition is itself an LLM judgment that creates the unit of analysis for all downstream metrics. Decomposition errors—missed claims, hallucinated claims, over-merged claims, or claims misaligned

with what the original segment asserts—propagate invisibly through the pipeline. A study of 100 segments (Section 5) shows 85% semantic overlap between LLM and independent human decomposition (95% CI: 77–91%), with the primary error mode being over-merging (11% of segments) rather than hallucination (3%). Systematic inter-annotator agreement analysis on decomposition quality across multiple annotators has not been conducted.

QA-grounding extension. Inspired by Q^2 (Honovich et al., 2021), a third scoring channel could detect passages that are topically relevant but answer a different sub-question than the claim. At three sequential LLM calls per (claim, passage) pair, it is cost-prohibitive at scale but represents a high-value future extension.

Scope and cost. All experiments use English-language datasets sourced primarily from Wikipedia (ASQA, ELI5) or similar curated corpora (HAGRID). Portability to other domains (scientific literature, legal text, code documentation) or non-English languages is untested. All evaluated and judge models are closed-source frontier systems accessed via API; results may not be fully reproducible as API weights change between versions. CiteScan requires approximately M decomposition calls, $N \times K$ entailment calls, and $N \times K$ embedding calls per sample (roughly 200 LLM + 200 embedding calls at $N \approx 10$, $K = 20$), making it suitable for periodic audits rather than continuous evaluation.

Ethics Statement

CiteScan proposes automated metrics for citation quality evaluation. We flag three ethical considerations.

Metric gaming (Goodhart's Law). Once a metric becomes an optimization target, systems may learn to satisfy it without improving the underlying property it measures. A RAG system trained to maximize OCA@1 might learn to always cite the passage with the highest embedding similarity to the query, regardless of genuine evidential support for the specific claim. We recommend CiteScan be used as a periodic diagnostic tool rather than a training signal, and that deployment decisions be accompanied by human spot-checks.

Citation failure labels carry reputational weight. UCR assigns citation failure labels to specific model outputs. Because our thresholds (τ_p, τ_f) are validated only on a single condition (ELI5-GPT5), automated labels could mischaracterize a system's behavior. We caution against using CiteScan scores for public model comparisons without accompanying human validation (see Limitations above).

Computational cost and environmental impact. A full CiteScan evaluation requires approximately 20,000+ LLM API calls per 100 samples. Responsible use requires weighing diagnostic value against energy consumption. The framework is designed for targeted audits, not continuous evaluation pipelines.

Dual-use potential. Detailed diagnostic information about citation failure modes could be used adversarially to train systems that produce convincing but fabricated citations. We consider this risk low: the information needed to produce convincing citations (entailment scores per passage) is already available to anyone with LLM API access, and CiteScan's contribution is in *detecting* these failures rather than enabling them. Nevertheless, we recommend responsible disclosure of model-specific vulnerability profiles.

References

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

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.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

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.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11346–11364.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9029–9057.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2463–2473.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*.

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.

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. 2025. Benchmarking LLM faithfulness in RAG with evolving leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 799–811.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. 2025. L-CiteEval: A suite for evaluating fidelity of long-context models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 5254–5277.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not faithfulness in RAG attributions. *arXiv preprint arXiv:2412.18004*.

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 North American Chapter of the Association for Computational Linguistics*, pages 547–565.

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*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.

A Evaluation Prompts

All three prompts are identical across providers; only the structured-output schema wrapper varies. Prompts are passed as the `system` message; the `user` message contains the data fields shown below. Brackets denote field substitution at runtime.

A.1 Claim Decomposition Prompt (system).

CLAIM DECOMPOSITION (SYSTEM)

```
You are a claim extraction specialist. Given a
USER QUERY, a SEGMENT from a chatbot response,
and the FULL RESPONSE for context, extract
the important factual claims from the SEGMENT
that are relevant to the user's query.

Rules:
1. Each claim must be ONE standalone factual
assertion understandable without the original
text.
2. Replace pronouns and references with the
specific entity they refer to.
3. Split compound sentences.
4. Skip headings, greetings, questions,
hedges, opinions, and meta-commentary.
5. Strip markdown, citation markers, and
URLs.
6. Preserve factual content accurately.
7. Extract only claims that directly address
or support the USER QUERY; when in doubt,
include.
8. Maximum 2 claims per segment; prefer claims
most central to answering the query.

Output JSON: {"claims": ["claim 1", ...]}.
If no relevant claims, return {"claims": []}.
```

User message: `USER QUERY: {query} / FULL RESPONSE: {response} / SEGMENT: {text}`

A.2 Entailment Prompt (system, Channel A).

ENTAILMENT JUDGE (SYSTEM, CHANNEL A)

```
You are a strict fact-verification judge.
Determine whether the CLAIM is true according
to the PASSAGE. The verdict is SUPPORTED if
and only if a reader with access only to
the passage would conclude the claim is true.
Otherwise: NOT_SUPPORTED.

Rules: Only use information explicitly stated
in the passage. Do not use your own knowledge
or inference. The passage must actually
make the claim true—merely mentioning related
topics is insufficient. If any part of the
claim adds, exaggerates, or distorts facts
beyond what the passage states, the verdict
is NOT_SUPPORTED.

Process:
1. Identify each distinct fact in the claim.
2. Find the passage sentence(s) that
establish it.
3. If every fact is established → SUPPORTED.
4. If any fact is unsupported →
NOT_SUPPORTED.

Output JSON: {"reasoning": "...", "verdict":
"SUPPORTED" or "NOT_SUPPORTED"}.
```

User message: `PASSAGE: {passage} / CLAIM: {claim}`

A.3 Relevance Classification Prompt (system, utility stage).

RELEVANCE CLASSIFIER (SYSTEM, UTILITY STAGE)

```
You are a relevance classifier for citation
quality evaluation. Classify how useful each
CLAIM is for answering the QUERY, using the
GOLD ANSWER as a reference compass (not an
exhaustive answer).

Labels:
 CORE — claim directly answers the query
 HELPFUL_DETAIL — useful supporting context
 OFF_TOPIC — does not help answer the query

Decision:
1. What is the user asking?
2. Does the claim address that need? No →
OFF_TOPIC.
3. Is it consistent with the gold answer's
direction?
   Contradicts → OFF_TOPIC.
   Directly answers → CORE.
   Supports or elaborates → HELPFUL_DETAIL.

Output JSON: {"label":
"CORE"/"HELPFUL_DETAIL"/"OFF_TOPIC",
"reasoning": "one sentence"}.
```

User message: `QUERY: {query} / GOLD ANSWER: {gold} / CLAIM: {claim}`

B RAG Response Generation Prompt

All evaluated models (GPT-5, Claude Sonnet 4.5, Gemini 2.5 Flash) receive the following system prompt for RAG response generation. The passage block is removed for PKR context-free generation (Appendix A, Section 3.5).

RAG Response Generation (System)

```
You are a helpful assistant. Answer the user's question using only the provided passages. For every claim or fact in your response, cite the relevant passage(s) using inline bracket notation (e.g., [1], [2], [1,3]). Only cite passages that directly support the claim you are making. Do not cite passages for general background or unsupported inference. If a claim cannot be supported by any provided passage, either omit it or state it without a citation.
```

User message: `Question: {query} / Passages: [1] {passage_1} ... [K] {passage_K}`

C Worked Example

We trace one complete evaluation through the CiteScan pipeline.

Input. Query: "Who wrote the book of 1 and 2 Thessalonians?" (ASQA sample `asqa_0028`). The retrieval set contains $K = 20$ passages including:

- c_0: Second Epistle to the Thessalonians (Wikipedia article opening)
- c_4: Authorship of the Pauline epistles (scholarly overview)
- c_5: First Epistle to the Thessalonians (Wikipedia article opening)

Step 1: Citation Parsing. The GPT-5 response produces three segments:

- s_1: "Both 1 Thess. and 2 Thess. are traditionally attributed to Paul..." $\mathcal{I}_1 = \{0, 5, 11, 14, 18\}$
- s_2: "Most modern scholars accept Paul as author of 1 Thess...." $\mathcal{I}_2 = \{4, 5\}$
- s_3: "Modern scholarship is divided on 2 Thess...." $\mathcal{I}_3 = \{0, 2, 4, 8, 13, 15, 19\}$

Step 2: Claim Decomposition. Five claims extracted (max 2 per segment):

d_1: "Both 1 and 2 Thessalonians are traditionally attributed to Paul the Apostle." ($\mathcal{I}_d = \{0, 5, 11, 14, 18\}$)

d_2: "The openings name Paul along with Silvanus and Timothy as senders." ($\mathcal{I}_d = \{0, 5, 11, 14, 18\}$)

d_3: "Most modern NT scholars accept Paul as author of 1 Thessalonians." ($\mathcal{I}_d = \{4, 5\}$)

d_4: "Modern scholarship is divided on whether Paul authored 2 Thessalonians." ($\mathcal{I}_d = \{0, 2, 4, 8, 13, 15, 19\}$)

d_5: "Many scholars consider 2 Thess. disputed or possibly pseudepigraphical." ($\mathcal{I}_d = \{0, 2, 4, 8, 13, 15, 19\}$)

Step 3: Scoring and Metrics. Each claim is scored against all 20 passages. Selected results:

Claim	OCA	CSM	Fab?	Rel.
d_1	0	-0.048	No	Core
d_2	0	-0.008	**Yes**	Helpful
d_3	1	$+0.481$	No	Core
d_4	1	$+0.007$	No	Core
d_5	1	$+0.463$	No	Core

Interpretation.

- d_1: OCA=0, negative CSM — an uncited passage scored higher than the cited set. Misattribution (not fabrication: some cited passage still exceeds τ_p).
- d_2: **Fabricated** — the cited passages score below $\tau_p = 0.3$ (the specific detail about "Silvanus and Timothy as senders" lacks explicit passage support) and no alternative exceeds $\tau_f = 0.6$.
- d_3: OCA=1, high CSM ($+0.481$) — c_4 (Pauline authorship article) strongly supports this claim with large margin over alternatives.
- d_4: OCA=1, near-zero CSM ($+0.007$) — "correct but lucky": the cited passage barely outscores alternatives. Multiple passages discuss Thessalonians authorship.
- d_5: OCA=1, high CSM ($+0.463$) — strong support with clear margin.

Sample-level aggregation. OCA@1 $= 3/5 = 0.60$; CSM $=$ mean$(\{-0.048, -0.008, 0.481, 0.007, 0.463\}) = 0.179$; CRR $= 5/5 = 1.0$ (all cited claims are Core or Helpful); MCR $= 0/5 = 0.0$; UCR $= 1/5 = 0.20$.

D Extended Experimental Details

Model identifiers and API dates. Model API calls were made between March–May 2026 using `gpt-5` (OpenAI, `reasoning_effort=none`), `claude-sonnet-4-5-20250514` (Anthropic), and `gemini-2.5-flash` (Google). Embedding model:

`text-embedding-3-large` (OpenAI, 3,072 dimensions). API model weights may change over time; we report identifiers as accessed during our evaluation window. Judge LLM calls use temperature 0 with structured JSON output (`max_tokens`=1,000); RAG response generation and PKR context-free generation use temperature 1.0 with `max_tokens`=4,096 across all three models. Rate limiting: 2,000 RPM with 50 concurrent requests per channel.

PKR entailment. PKR's param(d) check uses the same Channel A entailment mechanism: an LLM judges whether a context-free claim d' entails the RAG claim d (same prompt as Appendix A, A.2, with d' as passage and d as claim). This reuses the existing pipeline infrastructure without additional models.

Reproducibility. All evaluation code, prompts, and dataset preparation scripts are available at `[URL redacted for review]`. We will release all model-generated response files and claim-level score matrices at camera-ready to enable independent verification without re-running API calls. We evaluate all available samples per model-dataset pair ($n \approx 500$; ASQA-GPT5 $n = 496$ due to four empty-response instances). Samples were drawn sequentially from the first 500 entries of the dev split of each dataset as used by the ALCE evaluation framework (Gao et al., 2023).

Cost breakdown. CiteScan requires M decomposition calls, $N \times K$ entailment calls, and $N \times K$ embedding calls per sample, where M is the number of segments, N the number of claims, and K=20 the passage count. For a typical sample (M≈5, N≈10): 5 decomposition + 200 entailment + 200 embedding calls ≈ 205 LLM + 200 embedding API calls. At current pricing (GPT-5 input/output rates), approximately \$0.08–0.12 per sample. By comparison, ALCE's NLI-based evaluation uses a fixed classifier (<\$0.01/sample) but operates at sentence level without claim decomposition. CiteScan is designed for periodic diagnostic audits rather than continuous monitoring.

Retrieval quality context. Using Channel A entailment as a proxy, the fraction of passages with A=1 for at least one claim averages 0.72 on ASQA, 0.31 on ELI5, and 0.89 on HAGRID. This confirms that ELI5's higher UCR is partly attributable to lower passage relevance (the BM25+Contriever retriever produces tangentially relevant passages for open-ended explanatory queries), not solely to generation behavior.

E Robustness Analysis

This appendix presents the cross-judge validation, channel weight ablation, and threshold sensitivity experiments referenced in the Limitations section.

Judge	**UCR↓**	**OCA@1↑**
GPT-5 (self)	.371	.711
Claude Sonnet 4.5	.401	.699
Δ	+.030	−.012

Table 6: Cross-judge validation on 100 ELI5 samples (1,253 cited claims). Claude is marginally stricter (ΔUCR = +0.030) but directionally consistent.

α	**UCR↓**	**OCA@1↑**
0.00 (C-only)	.019	.642
0.25	.070	.699
0.50	**.401**	**.699**
0.75	.603	.699
1.00 (A-only)	.603	.964

Table 7: Channel weight ablation on 100 ELI5-GPT5 samples (1,253 cited claims). $\alpha = 0.5$ balances citation failure detection against false-positive suppression.

	τ_f				
τ_p	0.50	0.55	0.60	0.65	0.70
0.20	.058	.058	.058	.060	.061
0.25	.159	.159	.159	.160	.162
0.30	.327	.327	**.327**	.328	.333
0.35	.485	.485	.485	.487	.491
0.40	.526	.526	.526	.527	.532

Table 8: UCR on 50 ELI5-GPT5 samples across threshold pairs. Bold: operating point (τ_p=0.30, τ_f=0.60). τ_f has negligible effect (max Δ=0.006 per row); τ_p dominates sensitivity.