VERIFIED MISGUIDANCE: MEASURING STRUCTURAL CITATION FAILURES IN SEARCH-AUGMENTED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Users of search-augmented LLMs rely on citations as evidence that responses are grounded in real sources, and rarely verify the cited pages themselves. Millions of queries per day now pass through these systems, making citation quality a silent determinant of whether users are informed or misled—yet existing benchmarks address specific facets of citation quality, leaving the joint structure that determines citation trustworthiness unmeasured. We construct CITETRACE, a large-scale dataset that *traces the full citation chain from user query through retrieved source to generated answer*: 11,200 real-world queries from 28 communities paired with 112,000 responses from ten models across five providers, yielding 761,495 evaluable citation triples. We design a three-dimension evaluation framework that scores each citation on intent–purpose alignment, source suitability, and answer–source fidelity, using expert-validated predefined matrices and a five-level fidelity rubric; the framework applies to any system that produces citation-bearing responses. Applying this framework at scale, we identify a systematic pattern we call VERIFIED MISGUIDANCE (VM): models cite real, accessible sources yet fail along one or more dimensions, producing a fidelity–suitability trade-off in which faithful models select inappropriate sources and vice versa. Across our pool, 30.6% of citations distort their sources and 27.1% originate from domain-inappropriate sources; at the response level, up to 96% of users encounter at least one structurally misleading citation. Provider-level differences explain 88–96% of citation-quality variance, suggesting that source selection is governed more by factors beyond individual model capability than by the LLMs themselves. Together, CITETRACE and its evaluation framework provide the first resource for diagnosing structural citation failures in deployed search-augmented systems.

1 INTRODUCTION

Users of search-augmented LLMs rely on citations as evidence that responses are grounded in real sources (Ding et al., 2025; Narayanan Venkit et al., 2025), and rarely verify the cited pages themselves (Fogg et al., 2003; Liu et al., 2023). Millions of queries per day now pass through these systems (Miroyan et al., 2026), making citation quality a silent determinant of whether users are informed or misled. Search-augmented LLMs must judge query intent, select domain-appropriate sources through a retrieval pipeline whose division of labor between search backend and model remains opaque, and faithfully ground the answer in those sources. Each step in this multi-stage process can fail independently. On the same medical query, one model may cite a government health agency while another cites a personal wellness blog; both answers may look identical, but only one rests on a trustworthy source. The resulting failures are harder to detect than hallucination because the cited sources are real; we call this phenomenon **Verified Misguidance** (VM).

The tools available for auditing citations address specific facets of citation quality. Citation verification checks whether a claim is supported by its source (Gao et al., 2023; Rashkin et al., 2023); credibility rubrics such as the CRAAP test (Blakeslee, 2004) and Google's Your Money or Your Life (YMYL) classification (Google, 2025) rate source quality; generative search audits examine system-level source-selection patterns (Li & Sinnamon, 2024; Zhang et al., 2025). A source can pass a fidelity check and a quality rubric independently yet still be the wrong type of source for the user's question, and these tools may not flag the mismatch when applied separately.

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

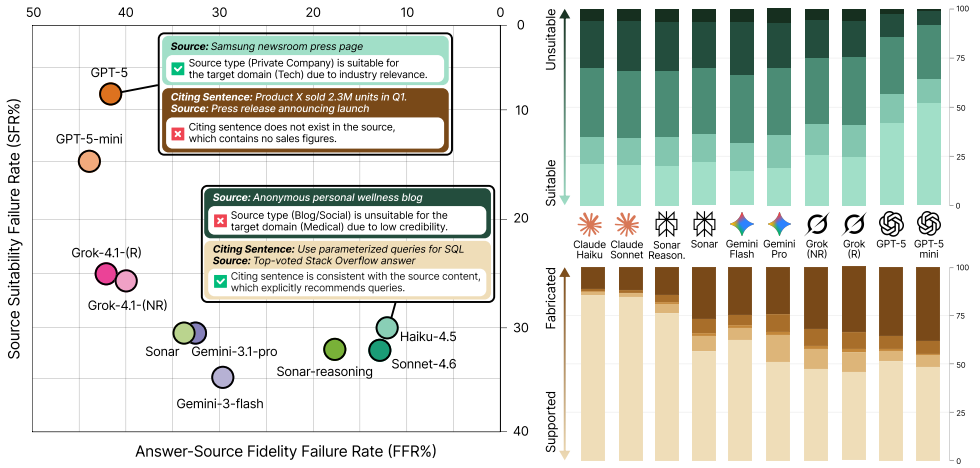

Figure 1: The fidelity–suitability trade-off across ten search-augmented LLMs. **Left**: each model is plotted by its Fidelity Failure Rate (FFR, x-axis) and Suitability Failure Rate (SFR, y-axis). `gpt-5` selects domain-appropriate sources (SFR 8.0%) but frequently fabricates their content (FFR 42.3%), while `claude-haiku` faithfully reflects its sources (FFR 12.3%) but draws heavily on unsuitable types (SFR 30.1%); the ideal corner remains unoccupied. Within each provider, models show similar fidelity–suitability profiles, while between-provider spread is large. **Right**: per-model distributions of source suitability (top, green) and answer–source fidelity (bottom, amber). Fidelity is sharply bipolar; most citations are either fully supported or fully fabricated, with few intermediate cases.

Diagnosing VM requires linking each citing sentence to the crawled content of its source across diverse query domains (Jin et al., 2026; Wu et al., 2025)—a combination that existing benchmarks provide only partially or for single domains. We construct CITETRACE to close this gap, sourcing cross-domain queries from 28 Stack Exchange communities where expert knowledge and commercial incentives coexist (Akerlof, 1970; Arrow, 1963), collecting search-augmented responses from ten LLMs across five providers under a neutral prompt that imposes no citation conventions, and crawling every cited URL to recover source content. The resulting dataset comprises **761,495** evaluable citation triples, each linking a user query to a model's citing sentence and the crawled content of its source. This three-way linkage enables, for the first time, joint analysis of whether the right source was selected and whether it was faithfully used.

To measure VM, we design a three-dimension evaluation framework: Intent–Purpose Alignment scores the fit between the query's information need and the source's communicative purpose; Source Suitability scores whether the source type is appropriate for the domain; and Answer–Source Fidelity scores whether the citing sentence faithfully reflects the source content. Each dimension captures a failure mode invisible to the other two: a source can be faithful yet unsuitable, suitable yet misaligned with the user's intent, or well-aligned yet unfaithful. All classifications rely on a single LLM judge (`gpt-4o-mini`) validated against human annotators (Appendix C). The framework is not tied to CITETRACE; it applies to any system that produces citation-bearing responses and can serve as a reusable instrument for citation quality evaluation beyond our setting.

Applying the framework to CITETRACE reveals that VM is pervasive across all ten models. Three patterns characterize the effect: (1) models that cite faithfully tend to select domain-inappropriate sources, and vice versa; a fidelity–suitability trade-off that no single model escapes; (2) 88–96% of citation-quality variance traces to the provider's search infrastructure rather than to model capability; and (3) these failures compound at the response level, exposing up to 96% of users to at least one structurally misleading citation (Figure 1). These patterns suggest that improving generation alone is insufficient; source selection at the retrieval stage warrants equal attention.

We make three contributions to the study of citation quality in search-augmented systems:

- **Dataset.** CITETRACE is the first large-scale resource that links real user queries, citation-bearing LLM responses, and crawled source content across 28 domains and ten models, enabling researchers to study citation quality jointly rather than one facet at a time (§2).

- **Evaluation.** Our three-dimension evaluation framework scores each citation on intent–purpose alignment, source suitability, and answer–source fidelity, three facets that existing tools evaluate separately. The framework applies to any system that produces citation-bearing responses (§3).
- **Findings.** Joint evaluation reveals a fidelity–suitability trade-off governed at the provider level, exposing up to 96% of users to at least one structurally misleading citation. These findings suggest that dedicated attention is needed not only at the generation stage but also at the stage where sources are retrieved and selected (§4).

CITETRACE is publicly released on HuggingFace at `https://huggingface.co/datasets/anon-citetrace/CiteTrace` with Croissant-compliant metadata (Akhtar et al., 2024), and the evaluation code is available at `https://anonymous.4open.science/r/CiteTrace2027`; licensing and reproducibility scope are documented in Appendix E.

2 CITETRACE: A LARGE-SCALE DATASET FOR CITATION QUALITY EVALUATION

Measuring search-augmented LLM citation quality requires three components: real user queries, citation-bearing model responses, and the content of each cited source. CITETRACE is a large-scale citation evaluation dataset that combines cross-domain queries, search-augmented responses from ten models, and the crawled content of their cited sources. We describe query sourcing (§2.1), response collection (§2.2), and source crawling (§2.3); full construction details appear in Appendix B.

2.1 SOURCING QUERIES

We source queries from Stack Exchange,[1] selecting 28 communities that satisfy three conditions jointly: expert knowledge is required to answer well, commercial actors are present and incentivized to shape perceptions, and the queries connect to substantive real-world decisions about safety, cost, or design. This combination is a key design choice; it targets precisely the settings where citation misguidance carries the highest cost and is hardest for users to detect. General-interest, entertainment, and single-correct-answer communities were excluded because they lack this commercial-bias dynamic. The 28 sites span Stack Exchange's six official categories (Appendix B, Figure 5), covering domains from medicine and law to programming and personal finance.

To ensure cross-domain comparability, we sample uniformly at 400 queries per site rather than proportionally, since proportional sampling would have been dominated by Stack Overflow alone. Combined with community-validation and recency filters, this yields a final dataset of 11,200 queries. Full selection criteria and per-site statistics appear in Appendix B.1.

2.2 COLLECTING RESPONSES AND CITATIONS

We query ten search-augmented LLMs from five providers (OpenAI, Anthropic, Google, xAI, and Perplexity), including standard and reasoning variant pairs where available. All models share a single neutral system prompt that specifies no citation format, count, or source preference, so that each model's citation behavior reflects its own defaults rather than prompt artifacts. We issue all 11,200 queries to each model within the same 15-day window to keep the web results comparable across models, producing 112,000 responses and 1,271,046 citation pairs; 8,069 responses (7.2%) contain no citations. Model identifiers, API configurations, and the full system prompt appear in Appendix B.2.

2.3 CRAWLING AND VERIFYING SOURCES

Unlike prior work that evaluates citations against a pre-retrieved corpus, we crawl every unique URL cited across the ten models directly, since citation quality can only be assessed against the actual content of the cited source. We retrieve 231,105 of 396,670 unique URLs (58.3%), covering 63.2% of all citation pairs, under `robots.txt` compliance.

Crawl failures are not uniform: they concentrate on forum and Q&A hosts due to bot blocking, which structurally under-represents community-based sources in the evaluable pool. Because these are

[1] https://archive.org/details/stackexchange_20251231

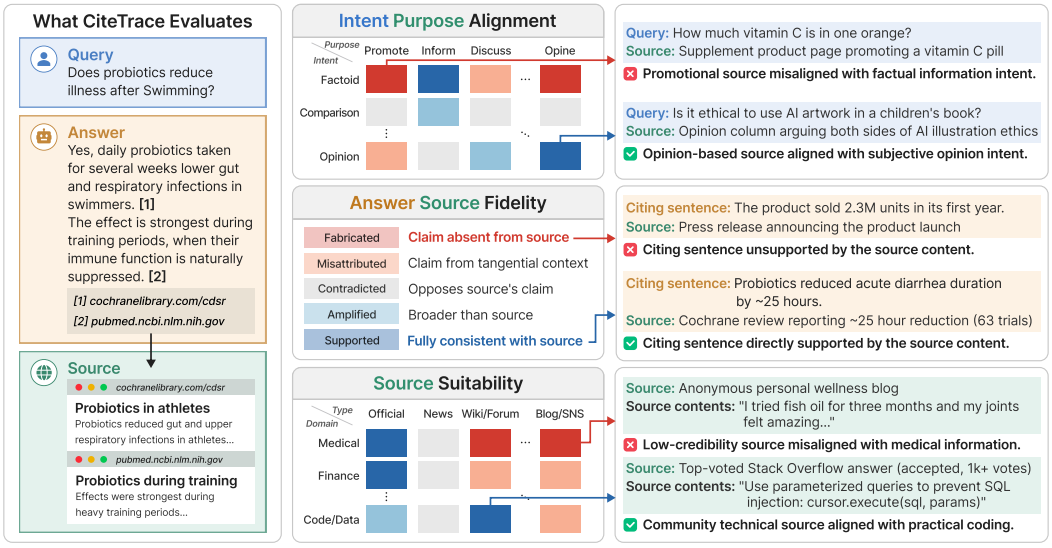

Figure 2: Overview of the three-dimension evaluation framework for diagnosing Verified Misguidance. **Left**: a user poses a query to a search-augmented LLM, which retrieves sources and generates a cited answer; the framework evaluates the resulting citation by jointly examining the query, the citing sentence, and the source content. **Center**: each citation is scored along three dimensions: Intent–Purpose Alignment (IPA) measures the fit between query intent and source purpose via a 5×6 matrix; Source Suitability (SS) measures whether the source type is appropriate for the domain via a 10×6 matrix; Answer–Source Fidelity (ASF) measures how faithfully the citing sentence reflects the source content on a five-level rubric. Each dimension yields a 1–5 score; scores ≤ 2 define dimension-level failure. **Right**: concrete pass/fail examples for each dimension: a factual query citing a promotional source (IPA failure), a fabricated claim absent from the source (ASF failure), and a medical query citing a personal blog (SS failure).

precisely the source types most likely to be problematic, our reported failure rates are conservative lower bounds on the true population values. After evaluability filtering, the final pool contains 761,495 citation triples across the ten models. The full crawl pipeline, failure breakdown, host-tier bias, and per-model evaluability rates appear in Appendix B.3.

3 An Evaluation Framework for Diagnosing Verified Misguidance

Having assembled the citation triples, the next question is how to evaluate them. Each triple comprises the query, the citing sentence, and the cited source content. A cited source can mislead the user even when it is real and the answer looks correct: its purpose may conflict with the user's intent, the source type may be inappropriate for the domain, or the answer may distort what the source actually says. To capture these three failure modes jointly, we propose a three-dimension evaluation framework (Figure 2): the fit between query intent and source purpose (§3.1), the suitability of the source type for the domain (§3.2), and the fidelity of the citing sentence to the source content (§3.3). The first two dimensions score each citation via expert-validated predefined matrices; the third uses a five-level fidelity rubric. The framework uses five classification tasks (query intent, source purpose, source domain, source type, and answer–source fidelity), all performed by a single LLM judge, whose reliability we validate against human annotators in Appendix C.

3.1 Alignment Between Query Intent and Source Purpose

A source can be factually accurate yet structurally misaligned with what the user needs; a promotional page cited for a factual query, or an opinion piece cited for a request seeking causal explanation. Factual verification alone cannot detect this, because the mismatch lies in the source's communicative function rather than its content. We classify each query by its intent (e.g., factoid, explanation,

instruction) (Bolotova et al., 2022) and each cited source by its communicative purpose (e.g., to promote, to inform, to discuss) (Biber et al., 2015; Sharoff, 2018), then score their alignment via a predefined 5×6 Intent–Purpose Alignment (IPA) Matrix on a 1–5 scale. Scores of 3–5 indicate functional alignment; scores of 1–2 flag structural conflict, where the source's incentive diverges from the user's need (Akerlof, 1970). We define citations with IPA score ≤ 2 as alignment failures and report their share as the Alignment Failure Rate (AFR). Full taxonomy, matrix design, and expert validation appear in Appendix C.1.

3.2 Source Suitability Across Domains and Types

Even when a source's purpose aligns with the user's intent, the source type may be inappropriate for the domain in question. The same personal blog may be acceptable for an everyday cooking question but structurally inappropriate for a medical query; yet existing source-quality frameworks evaluate sources independently of the domain they are cited for (Blakeslee, 2004). We score each source against a predefined 10×6 Source Suitability (SS) Matrix on a 1–5 scale that crosses substantive domain (e.g., medical, legal, finance) with publication type (e.g., official institution, paper/research, blog/social media). The matrix is anchored in the CRAAP test and Google's YMYL classification (Google, 2025), which define where source-type credibility requirements are highest. We define citations with SS score ≤ 2 as suitability failures and report their share as the Suitability Failure Rate (SFR). Full taxonomy, matrix design, and expert validation appear in Appendix C.2.

3.3 Fidelity of Answers to Their Sources

A source can be well-chosen yet still be misrepresented in the answer. Prior citation benchmarks reduce fidelity to a binary supported/not-attributable judgment (Gao et al., 2023; Rashkin et al., 2023), but the characteristic distortions of Verified Misguidance live between those endpoints: a claim may exist in the source yet be stripped of a critical qualifier, reversed in direction, or attached to tangential context that changes its meaning. We compare each citing sentence against the crawled source content and assign an Answer–Source Fidelity (ASF) score on a five-level rubric (Maynez et al., 2020) that explicitly separates these intermediate failure modes. We define citations with ASF score ≤ 2 as fidelity failures and report their share as the Fidelity Failure Rate (FFR). The five-level rubric and human-agreement statistics appear in Appendix C.3.

3.4 Validating the LLM Judge

The LLM judge is validated on 200 stratified samples per classification task, each independently labeled by three human annotators. Cohen's κ against majority-vote consensus ranges from 0.7960 to 0.8625 across the five tasks. Inter-annotator agreement among the three human annotators is Krippendorff's $\alpha \geq 0.809$, confirming that the taxonomy itself is stable and not merely an artifact of the judge. For the two matrix-based dimensions, expert panels of 10 domain specialists per domain independently rated each matrix cell; ICC(2,k) is 0.916 for the IPA Matrix and 0.958 (median) for the SS Matrix, with no cell deviating from expert consensus by two or more points. Together, these results support treating the judge's classifications as reliable inputs to the analyses in §4. Comparisons with two alternative judge models further support the choice of GPT-4o-mini, which achieves the highest agreement with human consensus on three of the five classification tasks. Full reliability statistics appear in Appendix C.

4 Structural Citation Failures in Search-Augmented LLMs

Applying the framework to the 761,495 evaluable citation triples in CiteTrace, we ask three questions: (1) how do failure rates differ across dimensions, and do they co-occur or trade off (§4.1); (2) what explains the variation: model capability or systematic provider-level differences (§4.2); and (3) how do citation-level failures compound into user-level exposure (§4.3). All reported rates are conservative lower bounds (Appendix D.6).

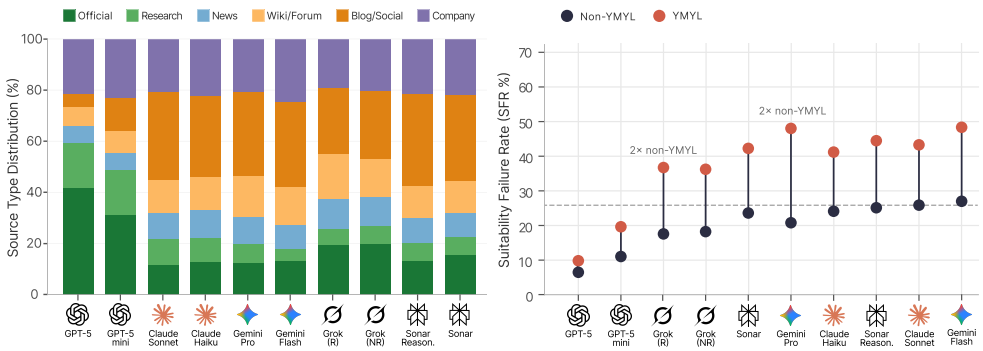

Figure 3: Source-type profiles and YMYL amplification. **Left**: source-type distribution per model, grouped by provider. Models within the same provider share similar profiles; the main variation across providers concentrates on the Official–Blog axis, while Company sources remain near-constant at 21%. **Right**: Suitability Failure Rate in YMYL domains (filled) versus non-YMYL domains (hollow) for each model. In YMYL domains, SFR nearly doubles the non-YMYL baseline across most models, highlighting that the risks of domain-inappropriate sourcing are substantially amplified precisely in the high-stakes queries where source suitability matters most.

4.1 The Three Dimensions Reveal Distinct Failure Patterns

We begin by examining how the three dimensions fail individually and whether their failures align or diverge across models. If citation quality were a single underlying construct, we would expect the three dimensions to co-vary; models that fail on one would tend to fail on others.

The source pool is structurally skewed before any dimension-level failure occurs. Blog/Social (29.2%) and Company (21.1%) sources together account for over half of all citations, while Official Institution (17.5%) and Research (8.8%) sources combined make up just over a quarter (Table 29). This skew is not model-specific but consistent across all ten models, suggesting it reflects the composition of sources that search backends surface rather than deliberate model choices (Appendix D.1).

Fidelity and suitability failures are prevalent but no single factor explains both. The three dimensions differ not only in failure rates but in what drives them. Fidelity failures are the most common (FFR 30.6%) and tend to be complete rather than partial: most citations are either fully supported or fully fabricated, with few intermediate distortions, suggesting that models make a near-binary choice to ground their claims in the source or ignore it entirely. The spread across models is large (12.3%–44.8%) and driven by model identity rather than query type (Appendix D.2). Source Suitability fails at a comparable rate (SFR 27.1%) but with a partly inverted model ranking, already hinting that fidelity and suitability are governed by different factors. Intent–Purpose Alignment has a failure rate about six times lower than fidelity failure (AFR 5.1%) and is shaped by query composition rather than model choice, reflecting the dominance of inform-purpose sources in the citation pool.

Models that cite faithfully tend to select unsuitable sources. The two main dimensions do not fail together but against each other (Figure 1). `claude-haiku` achieves the lowest FFR (12.3%) but ranks only 7th on suitability (SFR 30.1%), drawing 31.8% of its citations from Blog sources; `gpt-5` shows the reverse (FFR 42.3%, SFR 8.0%), drawing 41.8% from Official sources. Company sources account for a near-constant 21% across all models; the variation concentrates on the Official–Blog axis (Figure 3, Left). The inversion sharpens in YMYL domains, where SFR nearly doubles the non-YMYL baseline (Figure 3, Right). It also recurs along citation density: as citations per response grow, FFR drops (from 35% to 28%) while SFR rises (from 22% to 31%). No model occupies the ideal corner of both low FFR and low SFR. Models sharing a search backend exhibit similar source-type profiles, suggesting the trade-off reflects retrieval infrastructure rather than generation.

The three dimensions capture distinct failure modes. If the three dimensions captured a single underlying quality factor, citations that fail on one dimension would tend to fail on others. To examine their overlap, we define a Critical VM (CritVM) instance as a citation that fails on all three

dimensions simultaneously (i.e., IPA $\leq$2, ASF $\leq$2, and SS $\leq$2). Under an independence assumption, the expected CritVM rate is the product of the three marginal rates: $0.051 \times 0.306 \times 0.271 \approx 0.42\%$. The observed rate is 3,174 of 761,495 citations (0.42%), close to the independent-failure expectation. Single-dimension failures exceed three-way failures by at least a factor of five: ASF-only failures account for 22.2% of all citations, SS-only for 19.8%, and IPA-only for 2.1%, while the three-dimension intersection is 0.42%. This supports the view that the three dimensions capture structurally distinct failure modes, and that any single-dimension evaluation would miss failures captured by the other dimensions.

4.2 Provider-Level Differences, Not Model Capability, Explain Most Quality Variance

The fidelity–suitability trade-off suggests that citation quality may be shaped more by provider-level factors than by individual model capability. We test this with five within-provider comparisons that hold the search backend constant while varying model scale and reasoning capability, isolating how much room the generator has to affect each dimension.

Scaling up does not improve citation quality. Within-provider model pairs that differ in scale but share a search backend show marginal quality differences (Table 37). The Anthropic pair (`claude-sonnet` vs. `claude-haiku`) differs by less than 0.8 pp on FFR and 1.6 pp on SFR despite `claude-sonnet` issuing roughly twice as many citations per response. The Google pair reproduces the fidelity–suitability trade-off in miniature: `gemini-pro` improves SFR by 3.4 pp but worsens FFR by 3.1 pp. The largest within-provider gap (`gpt-5` vs. `gpt-5-mini`: $-$2.6 pp FFR, $-$6.3 pp SFR) remains far smaller than the between-provider gaps observed above, suggesting that scaling model capacity within the same provider has little effect on which sources the search backend returns.

Reasoning helps fidelity but not source selection. Paired reasoning and non-reasoning models sharing a search backend isolate the generation-side effect (Table 38). The xAI pair shows differences of at most 2.2 pp on all three dimensions. The Perplexity pair shows a sharper pattern: the reasoning model issues half as many citations and cuts FFR by 16 pp (18.0% vs. 34.1%), but SFR and AFR change by less than 2 pp. When reasoning helps, it appears to affect how the model uses a source rather than which sources are retrieved. Fidelity is the dimension most responsive to generation-side improvements; source selection appears largely outside the generator's control.

Provider identity accounts for most quality variance. A variance decomposition confirms the pattern (Table 36): provider effects account for 96% of alignment and suitability variance, leaving under 4% for within-provider model differences. Fidelity variance is also provider-dominated (88%), but the remaining 12% is attributable to model-level differences, consistent with the reasoning effect above. The asymmetry is structurally interpretable: alignment and suitability are determined by which sources the search backend retrieves, before the generator acts; fidelity additionally depends on how the generator uses those sources, leaving some room for model-level variation.

4.3 Citation Failures Compound into Widespread User Exposure

Citation-level failure rates understate the user-facing impact because failures compound across the multiple citations in each response. Even a modest per-citation failure rate translates into near-certain exposure at the response level.

Most responses contain at least one flawed citation. `claude-sonnet`'s FFR is only 13.1%, but with $\bar{n} = 12.7$ citations per response its response-level fidelity exposure reaches 62.9%—a 4.8-fold amplification. Response-level exposure ranges from 71.3% (`claude-haiku`, $\bar{n} = 5.3$) to 96.1% (`grok-reasoning`, $\bar{n} = 9.0$); even the best-performing model leaves nearly three in four responses with at least one structurally flawed citation (Figure 4, Left). The amplification tracks citation density: as $\bar{n}$ grows from the 1–5 bin to the 20+ bin, FFR falls (33.7% to 28.2%) while SFR rises (22.4% to 31.7%), so denser responses shift the failure mode from fidelity to suitability without reducing overall exposure (Figure 4, Right). The fidelity–suitability trade-off of §4.1 is not diluted but amplified at the response level: `claude-sonnet` achieves the third-lowest response-level FFR

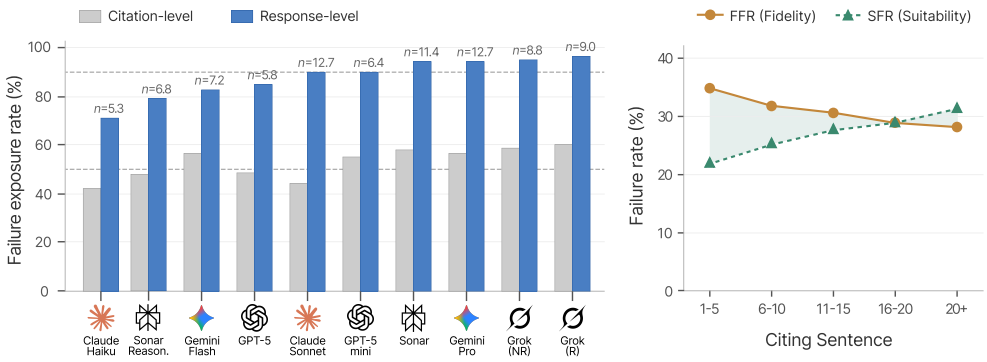

Figure 4: Response-level amplification of citation failures. A response is counted as exposed if at least one of its citations scores ≤ 2 on any of the three dimensions; the response-level rate is the share of such responses across all responses of that model. **Left**: citation-level failure rate (gray) versus response-level exposure (blue) for each model, sorted by response-level exposure. **Right**: fidelity and suitability failure rates by citation density bin. As citations per response grow, FFR decreases while SFR increases, reproducing the fidelity–suitability trade-off along the density axis.

yet the highest response-level SFR, confirming that no model offers reliable protection across both dimensions.

Some citations are unreachable before any dimension-level failure occurs. Compounding this further, a non-trivial share of citations points to sources that are entirely inaccessible at the time of use due to link rot or domain decommissioning: phantom-citation rates range from 2.1% to 15.5% across models, with Google's two models exceeding 14% (Table 42). One illustrative case involves a query about U.S. climate change projections, where a model cited the Fifth National Climate Assessment hosted at a government domain that had been decommissioned months prior. Had the source been accessible, it would have scored SS = 5, an optimal source selection degraded entirely by infrastructure decay rather than model error. These phantom citations represent a failure mode distinct from the structural mismatches we measure: the source was once real and authoritative, but is no longer verifiable by the user (Appendix D.6).

The same answer can rest on very different citation reliability. On a fish-oil vitamin-A query (YMYL Medical; Appendix D.8, Case 4), all ten models reached the same conclusion, yet Anthropic cited NIH while Perplexity and Google cited marketing blogs. For this query, the fraction of each model's evaluable citations with ASF $\leq$2 (case-specific FFR) was 0% for both Anthropic models versus 80–100% for Perplexity and Google models. This gap is invisible from the answer alone. On a VIX-futures contango query (YMYL Finance; Appendix D.8, Case 5), provider-mean suitability scores differed by approximately a factor of two. Together, these cases and the cross-dimension analysis in Appendix D.6 support evaluating the three dimensions jointly, as each captures failures missed by the others. Users cannot judge citation quality without examining each source individually, the very step that search-augmented responses are designed to spare them.

5 Related Work

Citation faithfulness and attribution verification. Whether a model's generated claim is supported by its cited source is a central question in citation quality research. Early work established the task through NLI-based and QA-based sentence-level adjudication (Bohnet et al., 2023; Gao et al., 2023; Rashkin et al., 2023), and verification was then scaled to full RAG pipelines through lightweight trained verifiers and automated evaluation frameworks (Es et al., 2024; Saad-Falcon et al., 2024; Tang et al., 2024); citation-specific tools further classify citation intent and apply full-text evidence reasoning (Haan, 2025; Nicholson et al., 2021). Multi-level faithfulness rubrics from abstractive summarization provide a finer-grained vocabulary for citation failures beyond binary support judgments (Kryscinski et al., 2020; Laban et al., 2022; Maynez et al., 2020; Min et al., 2023; Pagnoni et al., 2021; Wang et al., 2020). Empirical audits confirm that failures are pervasive: correct

answers can be unfaithful to their sources (Saxena et al., 2025; Wallat et al., 2025), up to 90% of medical LLM responses lack full source support (Wu et al., 2025), and fabricated or hallucinated citations add a further failure mode (Rao et al., 2026; Xu et al., 2026). All of these frameworks treat the A–S link as the sole criterion, leaving source appropriateness for the user's query out of scope.

Source quality and reliability assessment. A separate line of work evaluates the quality or reliability of sources independently of claim fidelity. Credibility research shows that users judge source trustworthiness through prominence, authority, and contextual fit (Fogg et al., 2003; Kakol et al., 2017; Rieh & Danielson, 2007; Sun et al., 2019), and evaluation rubrics such as the CRAAP test (Blakeslee, 2004) and SourceBench (Jin et al., 2026) operationalize these criteria at the document level; Google's YMYL classification (Google, 2025) and RAG reliability estimation (Hwang et al., 2025) extend this to high-stakes domains and retrieval-time signals. Theoretical grounding comes from the adverse-selection model (Akerlof, 1970) and information-asymmetry framework (Arrow, 1963), which explain how information quality degrades in markets for expert knowledge. These frameworks provide criteria for assessing source quality, but the same source can be appropriate for one query intent or domain and inappropriate for another.

Generative search engine audits. A growing body of work examines how search-augmented LLMs select and present sources at the system level. Source-selection audits show that generative engines disproportionately surface news and business sources (Li & Sinnamon, 2024) and differ from traditional search in coverage and citation patterns (Zhang et al., 2025); content producers have begun optimizing for these dynamics (Aggarwal et al., 2024; Kim et al., 2026). On the user side, citation count raises perceived credibility regardless of actual support (Ding et al., 2025; Miroyan et al., 2026). System-level evaluations identify structural failure modes (Narayanan Venkit et al., 2025), find that only 51.5% of generated sentences are fully supported by citations on average (Liu et al., 2023), and jointly assess credibility and groundedness (Vykopal et al., 2026). Each of these studies examines specific facets, including bias, preference, or verifiability, without connecting source selection to query intent or answer fidelity within a unified framework.

Positioning. Prior work measures A–S faithfulness (Bohnet et al., 2023; Es et al., 2024; Gao et al., 2023; Tang et al., 2024; Xu et al., 2025), source quality (Blakeslee, 2004; Hwang et al., 2025; Jin et al., 2026), or system-level source patterns (Li & Sinnamon, 2024; Miroyan et al., 2026; Zhang et al., 2025) in isolation; none model the conditional dependence between source type, query intent, and answer fidelity that drives the failure patterns we report. This fragmentation means that a system can score well on any single dimension while failing structurally on another, a gap that single-dimension benchmarks cannot detect by design. CITETRACE is the first dataset to measure all three simultaneously, enabling joint analysis of the fidelity–suitability trade-off, provider-level variance, and response-level exposure that prior benchmarks leave unmeasured.

6 CONCLUSION

We introduce CITETRACE, a large-scale dataset of 761,495 citation triples from ten search-augmented LLMs across 11,200 real-world queries, and a three-dimension evaluation framework for citation quality jointly across intent–purpose alignment, source suitability, and answer–source fidelity, aspects that existing benchmarks assess in isolation or not at all. VERIFIED MISGUIDANCE (VM) names the phenomenon in which search-augmented LLMs cite real, accessible sources that mislead through intent–purpose misalignment, domain-inappropriate sourcing, or distortion of source content. Our analyses demonstrate that models that cite faithfully tend to select unsuitable sources, while models that select suitable sources distort their content, a trade-off that appears to stem from factors beyond any single model's generation capability, and remains invisible to any single-dimension evaluation. Our findings suggest that improving generation alone may be insufficient to resolve citation quality, and that dedicated attention to source selection, at either the retrieval or generation stage, may be equally warranted. CITETRACE and our evaluation framework are released as a reusable resource; we hope they catalyze future efforts to improve source selection and establish citation quality as a first-class evaluation criterion for search-augmented systems. As these systems mediate access to information for millions of users daily, the structural failures we document are not edge cases but routine events whose scale and invisibility make them a concern for researchers and developers.

AI USE STATEMENT

We used LLMs to assist with experimental implementation and literature checking, and to polish the writing. The authors verified all references and reviewed all AI-assisted work. In the experiments, search-augmented LLMs generated responses and an LLM judge assigned classification labels, with human validation described in Appendix C. The authors take responsibility for the paper's final content and accuracy, including AI-assisted work.

ETHICS STATEMENT

This study uses publicly available Stack Exchange queries. Human annotators and expert raters participated as co-investigators or domain-expert collaborators without monetary compensation (Appendix C). We release source URLs and citing-sentence extracts rather than full crawled pages, with licensing and redistribution scope documented in Appendix E. Potential societal benefits and risks are discussed in detail in Appendix A.4.

REPRODUCIBILITY STATEMENT

We provide a static dataset snapshot and collection and analysis code for the Stack Exchange evaluation, with access information and reproducibility scope in Appendix E. Data collection and filtering procedures are described in Appendix B.1, Appendix B.2, and Appendix B.3, and evaluation and human validation procedures are further detailed in Appendix C.

REFERENCES

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 5–16, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671900. URL https://doi.org/10.1145/3637528.3671900.

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/1879431.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, Michael Kuchnik, Sylvain Lesage, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Hamidah Oderinwale, Pierre Ruyssen, Tim Santos, Rajat Shinde, Elena Simperl, Arjun Suresh, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Susheel Varma, Jos van der Velde, Steffen Vogler, Carole-Jean Wu, and Luyao Zhang. Croissant: A metadata format for ml-ready datasets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 82133–82148. Curran Associates, Inc., 2024. doi: 10.52202/079017-2610. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9547b09b722f2948ff3ddb5d86002bc0-Paper-Datasets_and_Benchmarks_Track.pdf.

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963. ISSN 00028282, 19447981. URL http://www.jstor.org/stable/1812044.

Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Heng Ji, Jong C. Park, and Rui Xia (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 122–131, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.15. URL https://aclanthology.org/2021.acl-demo.15/.

Douglas Biber, Jesse Egbert, and Mark Davies. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45, 2015.

Sarah Blakeslee. The craap test. *Loex Quarterly*, 31(3):4, 2004.

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, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models, 2023. URL https://arxiv.org/abs/2212.08037.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 1196–1207, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531926. URL https://doi.org/10.1145/3477495.3531926.

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792552. URL https://doi.org/10.1145/792550.792552.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Yifan Ding, Matthew Facciani, Ellen Joyce, Amrit Poudel, Sanmitra Bhattacharya, Balaji Veeramani, Sal Aguinaga, and Tim Weninger. Citations and trust in llm generated responses. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i22.34550. URL https://doi.org/10.1609/aaai.v39i22.34550.

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

B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, pp. 1–15, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137281. doi: 10.1145/997078.997097. URL https://doi.org/10.1145/997078.997097.

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

Google. Search quality evaluator guidelines. https://guidelines.raterhub.com/searchqualityevaluatorguidelines.pdf, 2025.

R. V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: evolution of structured data on the web. *Commun. ACM*, 59(2):44–51, January 2016. ISSN 0001-0782. doi: 10.1145/2844544. URL https://doi.org/10.1145/2844544.

Sebastian Haan. Semanticcite: Citation verification with ai-powered full-text analysis and evidence-based reasoning, 2025. URL https://arxiv.org/abs/2511.16198.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34279–34303, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1738. URL https://aclanthology.org/2025.emnlp-main.1738/.

IAB Tech Lab. Content taxonomy 3.0. IAB Tech Lab Standard, 2022. URL https://iabtechlab.com/standards/content-taxonomy/.

Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Information Retrieval Series*. Springer, Dordrecht, 2005. ISBN 978-1-4020-3851-8. doi: 10.1007/1-4020-3851-8.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, May 2008. ISSN 0306-4573. doi: 10.1016/j.ipm.2007.07.015. URL https://doi.org/10.1016/j.ipm.2007.07.015.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Hexi Jin, Stephen Liu, Yuheng Li, Simran Malik, and Yiying Zhang. Sourcebench: Can ai answers reference quality web sources?, 2026. URL https://arxiv.org/abs/2602.16942.

Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 53(5): 1043–1061, 2017. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2017.04.003. URL https://www.sciencedirect.com/science/article/pii/S0306457316306471.

Sunghwan Kim, Wooseok Jeong, Serin Kim, Sangam Lee, and Dongha Lee. Sageo arena: A realistic environment for evaluating search-augmented generative engine optimization. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2342–2353, 2026.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344. Curran Associates, Inc., 2024. doi: 10.52202/079017-3342. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be2e1b68b44f2419e19f6c35a1b8cf35-Paper-Datasets_and_Benchmarks_Track.pdf.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, June 2016. doi: 10.1016/j.jcm.2016.02.012.

M. Koster, G. Illyes, H. Zeller, and L. Sassman. Robots Exclusion Protocol. RFC 9309, September 2022. URL https://www.rfc-editor.org/info/rfc9309/.

Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011. URL https://api.semanticscholar.org/CorpusID:59901023.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750/.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

Alice Li and Luanne Sinnamon. Generative ai search engines as arbiters of public knowledge: An audit of bias and authority. *Proceedings of the Association for Information Science and Technology*, 61(1):205–217, 2024. doi: https://doi.org/10.1002/pra2.1021. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.1021.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios Angelopoulos, trevor darrell, Narges Norouzi, and Joseph E Gonzalez. Search arena: Analyzing search-augmented llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 41109–41146, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4476dd7320e0eba63961990d73525064-Paper-Conference.pdf.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. Search engines in the ai era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in llm-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 1325–1340, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732089. URL https://doi.org/10.1145/3715275.3732089.

Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3): 882–898, 11 2021. ISSN 2641-3337. doi: 10.1162/qss_a_00146. URL https://doi.org/10.1162/qss_a_00146.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812–4829, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.383. URL https://aclanthology.org/2021.naacl-main.383/.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. Citeme: Can language models accurately cite scientific claims? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 7847–7877. Curran Associates, Inc., 2024. doi: 10.52202/079017-0252. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ef47f7b768e1a012e3d995ac8d8fac7-Paper-Datasets_and_Benchmarks_Track.pdf.

Delip Rao, Eric Wong, and Chris Callison-Burch. Detecting and correcting reference hallucinations in commercial llms and deep research agents, 2026. URL https://arxiv.org/abs/2604.03173.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, December 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Soo Young Rieh and David R. Danielson. Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, 41(1):307–364, 2007. doi: 10.1002/aris.2007.1440410114.

Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pp. 13–19, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 158113844X. doi: 10.1145/988672.988675. URL https://doi.org/10.1145/988672.988675.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Yash Saxena, Raviteja Bommireddy, Ankur Padia, and Manas Gaur. Generation-time vs. post-hoc citation: A holistic evaluation of llm attribution, 2025. URL https://arxiv.org/abs/2509.21557.

Serge Sharoff. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95, 2018.

Yalin Sun, Yan Zhang, Jacek Gwizdka, and Ciaran B Trace. Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *J Med Internet Res*, 21(5):e12522, May 2019. ISSN 1438-8871. doi: 10.2196/12522. URL https://www.jmir.org/2019/5/e12522/.

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, 1990.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

Minaoar Hossain Tanzil, Shaiful Chowdhury, Somayeh Modaberi, Gias Uddin, and Hadi Hemmati. A systematic mapping study of crowd knowledge enhanced software engineering research using stack overflow. *Journal of Systems and Software*, 226:112405, 2025. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2025.112405. URL https://www.sciencedirect.com/science/article/pii/S0164121225000731.

Robert S Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3):178–194, 1968. doi: 10.5860/crl_29_03_178.

Ivan Vykopal, Mat'uš Pikuliak, Simon Ostermann, and Marian Simko. Assessing web search credibility and response groundedness in chat assistants. In Vera Demberg, Kentaro Inui, and Llu'is Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2539–2560, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.115. URL https://aclanthology.org/2026.eacl-long.115/.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, pp. 22–32, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618. doi: 10.1145/3731120.3744592. URL https://doi.org/10.1145/3731120.3744592.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450/.

Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. An automated framework for assessing how well llms cite relevant medical references. *Nature Communications*, 16(1):3615, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58551-6. URL https://doi.org/10.1038/s41467-025-58551-6.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. ALiiCE: Evaluating positional fine-grained citation generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 545–561, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.23. URL https://aclanthology.org/2025.naacl-long.23/.

Zuyao Xu, Yuqi Qiu, Lu Sun, Fasheng Miao, Fubin Wu, Xiang Li, Xinyi Wang, Haozhe Lu, Zhengze Zhang, Yuxin Hu, et al. Ghostcite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718*, 2026.

Peixian Zhang, Qiming Ye, Zifan Peng, Kiran Garimella, and Gareth Tyson. Source coverage and citation bias in llm-based vs. traditional search engines, 2025. URL https://arxiv.org/abs/2512.09483.

Appendices

A Discussions: Scope, Limitations, and Broader Impact **17**
A.1 Scope and Operational Assumptions 17
A.2 Limitations 17
A.3 Future Directions 18
A.4 Broader Impact 19
A.5 Comparison to Adjacent Failure Modes 19

B CITETRACE: Dataset Construction **20**
B.1 Sourcing Real-World Queries from Stack Exchange Communities 20
B.2 Collecting Search-Augmented Responses and Extracting Citations 22
B.3 Crawling and Verifying Source Content 24

C Three-Dimension Evaluation Framework **27**
C.1 Query–Source Alignment 28
C.2 Source Suitability 30
C.3 Answer–Source Fidelity 37
C.4 Cross-Dimension Integration and Robustness 38

D The VM Effect: Detailed Empirical Results **48**
D.1 A Structurally Biased Source Pool 48
D.2 Aggregate Failure Rates Across Three Dimensions 48
D.3 The Fidelity–Suitability Trade-off Across Models 51
D.4 How Provider, Scale, and Reasoning Shape Citation Quality 53
D.5 Response-Level Failure Exposure 55
D.6 Robustness and Lower Bounds 56
D.7 External Validation on Search Arena 58
D.8 Qualitative Failure Analysis 58

E Data Release and Reproducibility **65**
E.1 Data Access and Licensing 65
E.2 Schema and Field Documentation 65
E.3 Reproducibility Notes 67

A Discussion: Scope, Limitations, and Broader Impact

This appendix articulates the scope of evaluative claims our benchmark supports (§A.1), identifies five categories of limitations and their downstream effects (§A.2), proposes four research directions through which the framework can be extended (§A.3), discusses societal and engineering implications (§A.4), and positions VM relative to adjacent failure modes (§A.5).

A.1 Scope and Operational Assumptions

Construct. Verified Misguidance (VM) refers to the structural condition under which a citation-bearing answer may mislead users despite the cited source being real and accessible, even when faithfully cited, a condition distinct from hallucination (Ji et al., 2023). We operationalize this condition using three distinct dimensions: query intent–source purpose alignment, answer–source fidelity, and source-type suitability for the domain. Simultaneous failure on all three dimensions constitutes Critical VM (CritVM). The three dimensions capture distinct failure modes; as Appendix D.6 shows, single-dimension failures exceed failures on all three dimensions by at least a factor of five, so a single-dimension evaluation framework would miss the structural patterns we report.

Supported claims. CiteTrace supports three claim types: *existence*, that citation triples failing all three dimensions are measurable and non-degenerate within our setting (28 communities, 10 models, English, March–April 2026); *comparison*, that ranking-level differences across models, providers, and domains are robust (Appendix D.4, D.6); and *diagnosis*, identifying which axis fails individually and how the three co-fail (Appendix D.3–D.4). We do *not* support claims of permanent provider quality, generalization to non-English settings, or causal isolation of retrieval versus generation.

Measurement assumptions. Five assumptions support our measurements; each is paired with a corresponding limitation in §A.2.

A1 Citation-marker semantics. We interpret each provider's citation marker as asserting source–answer grounding, consistent with all five providers' API documentation (Appendix B.2).

A2 Predefined-matrix validity. IPA and SS matrices are expert-designed instruments, supported by expert validation (Table 25) and failure-threshold sensitivity analysis (Appendix D.6).

A3 LLM-judge reliability. GPT-4o-mini's per-task κ (0.7960–0.8625) meets the Landis–Koch substantial-agreement threshold (Landis & Koch, 1977); three-annotator human validation (Krippendorff $\alpha \geq 0.809$ (Krippendorff, 2011)) supports taxonomy stability (Appendix C).

A4 Query representativeness. Stack Exchange Q&A represents one dimension of real-world information needs; we make no claim about conversational or multi-turn search.

A5 Temporal snapshot. Results reflect a 15-day window; we do not claim trends outside this window.

A.2 Limitations

We acknowledge five categories of limitations that bound the interpretation of our findings, pairing each with its downstream effect.

Crawl-failure coverage bias. For 36.8% of citations, the cited source could not be crawled, with failures concentrated on Forum/Q&A (67.9%) and Social/Blog (55.0%) hosts due to Cloudflare bot blocking (Appendix D.6). The downstream metrics (AFR, SFR, FFR) are computed only over successfully crawled citations and therefore exclude these failures from the denominator. Future work could supplement live crawling with archival snapshots (e.g., the Internet Archive) for bot-blocked URLs.

LLM-as-Judge dependence. All taxonomy classifications and fidelity adjudication share a single LLM dependency (GPT-4o-mini), which may propagate systematic biases into downstream metrics in ways not fully captured by aggregate κ. This single-judge design follows the established methodology of comparable frameworks (ALCE (Gao et al., 2023) validates against a single NLI model, RAGAS (Es et al., 2024) prompts a single LLM, and ARES (Saad-Falcon et al., 2024) trains one

judge per dimension). Our judge achieves Cohen's κ of 0.7960–0.8625 against human majority-vote consensus across the five classification tasks. Nonetheless, absolute values may shift under judge replacement; we report *relative* cross-model and cross-provider patterns as the primary claims and encourage replication with an alternative judge family.

Predefined-matrix subjectivity. The IPA and SS matrices are expert-designed instruments rather than data-derived scores. ICC (Koo & Li, 2016) measures agreement among expert ratings, while failure-threshold sensitivity analyses show Kendall $\tau \geq 0.82$ for model-level CritVM rankings in five of seven settings. Nevertheless, reasonable experts may assign different scores to specific cells. A more ambitious extension would elicit cell values from a structured panel of 20–50 experts per domain, allowing publication of a probabilistic version of each matrix.

Linguistic and cultural scope. CITETRACE's 11,200 queries are drawn from English-language Stack Exchange. Information-asymmetry dynamics may differ in non-English contexts where the commercial-source landscape, regulatory environment, and platform ecosystems differ; for example, in markets with prominent state-affiliated media or single-platform Q&A communities (Naver Knowledge-iN, Zhihu), our SP/ST taxonomies may need refinement. We view cross-cultural extension (§A.3) as a high-priority direction.

Temporal snapshot and causal attribution. Each provider's search tool issues live web queries, so identical prompts may return different sources at different times. We minimize variation through a 15-day collection window, with all axes within ±5 percentage points across the window's two halves (Appendix D.6), but full reproducibility is not attainable. Our framework also characterizes *where* VM concentrates but does not isolate *which pipeline stage* (retrieval, generation, or interaction) produces it. Controlled experiments with fixed retrieval or fixed generation would enable sharper causal decomposition; this is currently blocked by the closed-source nature of commercial retrieval pipelines.

A.3 FUTURE DIRECTIONS

The limitations above suggest four concrete research directions.

Multilingual and cross-cultural extension. Extending CITETRACE to non-English Stack Exchange and to non-Stack-Exchange platforms (Zhihu, Naver Knowledge-iN, Reddit) would test whether VM concentrates on the same domain–source-type combinations across cultures and whether the 88–96% provider variance dominance holds when search backends are tuned to non-English markets.

Longitudinal observation. A quarterly variant (CITETRACE-Longitudinal) re-issuing the same 11,200 queries would track whether absolute FFR/SFR/AFR levels are improving, whether cross-provider trade-offs persist, and whether YMYL-domain failures narrow under retrieval-policy updates. The marginal collection cost is modest, and the result is a continuous monitoring signal that single-shot benchmarks cannot provide.

Causal decomposition through controlled retrieval. A controlled experiment with the retrieved-context set held fixed while the generator is varied (and vice versa) would isolate the contribution of each pipeline stage, moving from descriptive characterization (where does VM concentrate?) to mechanistic understanding (which pipeline stage produces it?).

Mitigation-oriented benchmarking. Three concrete mitigation paths follow from our findings: *intent-aware retrieval reranking* (inverting IPA matrix as a scoring function), *domain-aware source filtering for YMYL* (given the 2.3-fold Fisher OR), and *citation-quality nudging at generation time* (since reasoning models reduce FFR but not SFR). We release the framework, matrices, and per-citation labels to support replication of these mitigation experiments.

A.4 BROADER IMPACT

Societal implications: misguidance at scale. Search-augmented LLMs increasingly mediate access to information in high-stakes decision-making domains. Hundreds of millions of users now consult these systems in lieu of traditional search engines (Miroyan et al., 2026), treating the presence of a citation as a sufficient guarantee of accuracy (Ding et al., 2025). Our work demonstrates empirically that this trust is not always warranted: 30.6% of citations distort their sources, 27.1% originate from domain-inappropriate sources, and at the response level up to 96% of users encounter at least one such citation. Unlike a hallucinated fact, a citation-bearing answer that is faithful to a structurally inappropriate source presents no surface signal that distinguishes it from a trustworthy one; at scale, even single-digit failure rates translate into millions of misguidance events per day in the YMYL domains where misguidance carries the highest cost.

Implications for AI evaluation and system design. Our cross-dimension analysis (Appendix C.4) shows that a system can score well on any single citation-quality dimension while failing structurally on another: ALCE-style citation precision (Gao et al., 2023) and RAGAS-style faithfulness (Es et al., 2024) target answer-source fidelity, while CRAAP- and SourceBench-style source quality (Blakeslee, 2004; Jin et al., 2026) target source suitability, and neither captures their interaction. CITETRACE is designed to support multi-dimension structural evaluation in which a system's citation behavior is reported as a profile across IPA, SS, and ASF rather than a scalar. For practitioners, our finding that 88–96% of citation-quality variance is provider-level (Appendix D.4) implies that retrieval-backend choice matters more than generator choice for citation quality, and that fidelity (Answer–Source Fidelity) and source-selection (Query–Source Alignment and Source Suitability) require separate engineering efforts.

Responsibility in model comparison. Our work reports systematic differences across providers, including ranking reversals between fidelity and source suitability. These differences reflect specific model versions during a specific collection period and are *not* permanent properties of any provider; search-augmented systems evolve on weekly timescales, and the rankings may invert within a quarter. We release model identifiers and collection dates alongside all results to discourage citation of our rankings as static characterizations.

Ethics, safeguards, and misuse risk. All queries derive from the Stack Exchange Data Dump (CC BY-SA 4.0), and we comply with attribution requirements. Source crawling respects `robots.txt` and rate limits (Appendix B.3); the public release contains URLs and metadata only, not source bodies (Appendix E.1); no human-subjects data are collected. In principle, content producers could reverse-engineer IPA matrix to boost their pages' search-result visibility. We assess this risk as low because IPA matrix is a diagnostic tool, not a retrieval signal used by any production system.

A.5 COMPARISON TO ADJACENT FAILURE MODES

VM occupies a position adjacent to several phenomena studied in the LLM evaluation literature, and distinguishing it from each is essential for interpreting our findings.

Hallucination. Hallucination (Ji et al., 2023) concerns claims that are factually false or unsupported by any source. VM is structurally distinct: by construction, the cited source in a VM instance is real and accessible, and the cited claim need not be false. The structural failure lies in alignment of source purpose to query intent (Intent–Purpose Alignment) or in source-type suitability for the domain (Source Suitability), not in factual accuracy per se. Empirically, ASF1 (Fabricated) cases (closest to classical hallucination) account for 24.5% of citations, while structural failures on IPA or SS without fabrication add roughly another 20 percentage points.

Citation-precision benchmarks. ALCE (Gao et al., 2023), AutoAIS (Bohnet et al., 2023), and related work measure whether each cited claim is supported by the cited source. This corresponds to a portion of our Answer–Source Fidelity (ASF), but treats the suitability of the cited source as exogenous: if the source supports the claim, the citation is counted as correct, regardless of whether the source is structurally suitable for the user's intent. Our cross-dimension trade-off (Anthropic high-ASF/low-SS versus OpenAI low-ASF/high-SS) shows precisely why this is insufficient.

Source-quality benchmarks. SourceBench (Jin et al., 2026), the CRAAP test (Blakeslee, 2004), and related work evaluate the absolute quality of retrieved sources with a universal rubric, scoring each source independently of how its type interacts with its content domain. Our SS matrix (Appendix C.2) explicitly encodes this interaction: the same Wiki/Forum type is rated SS 5 when the source falls in Code/Data and SS 1 when it falls in Medical. A benchmark that ignores the domain–type interaction would conflate these cases and miss the YMYL-specific failure pattern that drives much of our reported SFR (38.3% in YMYL versus 21.1% outside).

Information asymmetry and trust calibration. Our work inherits the information-asymmetry frame of Akerlof (1970); Arrow (1963) but operationalizes it for the search-augmented LLM setting, where the LLM rather than the user is the proximate consumer of search results, and the user receives a summary that hides the structural mismatch from view. Recent trust-calibration research (Miroyan et al., 2026) shows that citation count raises user trust regardless of whether citations support the claim; our work contributes the system-side measurement that complements this user-side finding, quantifying the gap between surface citation behavior and the three structural quality axes.

B CITETRACE: DATASET CONSTRUCTION

This appendix details the dataset construction procedure summarized in Section 2. We describe how we source queries from Stack Exchange (§B.1), collect search-augmented responses and extract citations (§B.2), and crawl and filter cited sources (§B.3).

B.1 SOURCING REAL-WORLD QUERIES FROM STACK EXCHANGE COMMUNITIES

Site selection. Stack Exchange Network (Tanzil et al., 2025), based on its December 31, 2025 snapshot, hosts $182+$ communities. We narrow this pool to 28 sites through a four-stage selection process designed to retain communities where commercial information asymmetry is structurally possible. *Stage 0 (candidate pool,* $182+$ *down to* 48*)* excludes communities that are unsuitable for studying commercial citation bias on substantive grounds: single-correct-answer domains (e.g., math, codegolf), fiction and entertainment sites lacking decision-making contexts (gaming, anime, movies), language-correction sites without commercial actors (english, ell), platform-specific tools dominated by a single ecosystem (emacs, tex, blender), religious-doctrine sites whose claims resist factual verification (christianity, islam), sites lacking a structural commercial-bias mechanism (lifehacks, history), and beta sites with fewer than ten thousand valid questions. The full exclusion criteria appear in Table 1. *Stages 1–3 (*48 *down to* 28*)* apply five sequential gates summarized in Table 2: requirement of expert knowledge (C1), presence of commercial actors with incentive to distort answers (C2), connection to substantive decisions (C3), distinctness of the commercial-bias mechanism across selected sites (C4), and data sufficiency after quality filtering (C5). Criterion C1 is satisfied by all 48 Stage 0 candidates by construction; C2–C3 remove three sites with absent or unverified commercial actors (interpersonal, philosophy, ux); C4 deduplicates eight sites that share an existing site's bias mechanism; and C5 removes nine sites whose post-filter pool falls below 400 queries, ensuring sufficient data for our per-site sampling target. The selection gates were applied through consensus discussion rather than independent ratings, with disagreements and borderline cases resolved through discussion. The remaining 28 sites map directly to Stack Exchange's official six-category taxonomy: Technology 5, Science 7, Life & Arts 7, Culture & Recreation 6, Professional 2, Business 1. This category distribution emerged from the selected sites rather than a predefined allocation. We follow PRISM's (Kirk et al., 2024) approach of adopting an external taxonomy (UN subregion in their case) to avoid imposing our own categorization.

Quality filtering and sampling. A five-step filter applied to all 28 sites yields a final pool of $471{,}876$ eligible queries from $25{,}483{,}555$ raw posts, a 98.1% reduction summarized stage-by-stage in Table 3. The single largest reduction is the community-validation gate (Score ≥ 5, 90.96% drop), which removes unanswered or low-quality posts; the recency gate ($\geq 2018{-}01{-}01$, 78.91% drop) further restricts the pool to queries posed within the search-augmented LLM era. We then apply uniform per-site sampling of 400 queries to obtain the final $11{,}200$-query dataset (28×400). Uniform sampling is essential because Stack Overflow alone accounts for 81.8% of the eligible pool ($385{,}770/471{,}876$); proportional sampling would make cross-site comparisons impossible.

Table 1: Stage 0 exclusion criteria for Stack Exchange site selection. Sites meeting any single criterion are removed from consideration before the substantive gates (C2–C5) are applied.

Exclusion criterion	**Rationale**	**Excluded sites (examples)**
Single-correct-answer domain	Commercial bias does not arise structurally	math, puzzles, codegolf
Fiction & entertainment	No decision-making context	gaming, anime, scifi, movies
Language correction	Commercial actors absent	english, ell
Platform-specific tools	Single-tool ecosystem yields trivial bias	emacs, tex, blender, drupal
Religious doctrine	Belief claims resist verification	christianity, islam, judaism
Absent commercial structure	No commercial-bias mechanism	lifehacks, history
Insufficient scale ($<$ 10K Qs)	Beta sites, inadequate sample size	various beta sites

Table 2: Selection gates C1–C5 applied in Stages 1–3. C1 is automatically satisfied by all Stage 0 survivors; C2–C5 are the substantive gates that reduce the candidate pool from 48 to 28 sites.

#	**Criterion**	**Decision question**
C1	Expertise required	Does answering require domain expertise?
C2	Commercial actor present	Are there commercial stakeholders incentivized to distort answers?
C3	Decision-making context	Does the question relate to substantive decisions (safety, cost, design)?
C4	Bias-mechanism uniqueness	Is the commercial-bias mechanism distinct from other selected sites?
C5	Data sufficiency	Are at least 400 queries available after quality filtering?

Table 3: Five-step quality filter funnel applied to all 28 sites combined. The community-validation gate (Step 1) and recency gate (Step 4) account for 99% of the total reduction; per-site uniform sampling (Step 6) produces the final 11,200-query dataset.

Stage	**Filter**	**Remaining**	**Stage drop %**
0	Total (28 sites combined)	25,483,555	–
1	Score ≥ 5	2,303,787	90.96
2	Title 15–150 chars	2,303,628	0.01
3	Body ≥ 50 chars (HTML stripped)	2,303,487	0.01
4	Posted $\geq 2018{-}01{-}01$	485,816	78.91
5	Intent classified (drop `unclear`)	471,876	2.87
6	Per-site uniform sample of 400	**11,200**	–

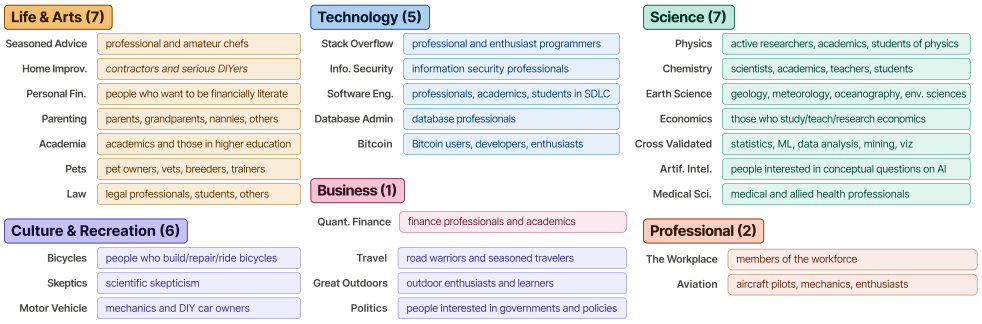

Figure 5: Composition of the CITETRACE query set. The 28 Stack Exchange communities are grouped by Stack Exchange's six official categories. Each community is annotated with its audience and domain. Communities were selected to satisfy three conditions: expert knowledge is required, commercial actors are present, and queries involve substantive decisions about safety, cost, or design.

Per-site composition. The selected 28 sites span Stack Exchange's six categories with the distribution Technology 5, Science 7, Life & Arts 7, Culture & Recreation 6, Professional 2, Business 1. Table 4 reports for each site the official category, audience description (verbatim from the Stack Exchange Sites API), pool size after Step 5 of the quality filter, sample size (400 uniform), and mean posting year. The mean year across the full pool is 2020.27, indicating that the dataset reflects the period when search-augmented LLMs entered mainstream use.

B.2 COLLECTING SEARCH-AUGMENTED RESPONSES AND EXTRACTING CITATIONS

Model selection. We select ten search-augmented LLMs from five major providers, including standard/reasoning variant pairs to enable controlled comparisons of reasoning effects on citation behavior. GPT-5 and GPT-5 mini (OpenAI), Claude Sonnet 4.6 and Claude Haiku 4.5 (Anthropic), Gemini 3 Flash and Gemini 3.1 Pro (Google), Grok 4.1 Fast-Non-Reasoning and Grok-4.1 Fast-Reasoning (xAI), and Sonar and Sonar Reasoning Pro (Perplexity). Selection criteria are: (i) provider diversity to capture distinct retrieval-backend behaviors, (ii) built-in search tool support via the production API, (iii) non-deprecated availability throughout the collection window, and (iv) the existence of at least one standard/reasoning pair within each provider that supports such pairing. Table 5 summarizes per-model API parameters, search tools, and collection windows.

Prompt design. We use a single system prompt across all ten models to isolate citation behavior specific to each model from variance introduced by the prompt. The prompt is designed to intervene as little as possible, providing no specification of citation format, count, or preference for particular source types so that each model's natural citation behavior can surface. At the same time, it preserves a realistic query format by passing only the Stack Exchange post title as the user query, mimicking the short questions users typically ask rather than inputs that have been artificially expanded. The full system prompt appears in Figure 6.

Response collection. We dispatch each of the 11,200 queries to each of the 10 models, yielding $11{,}200 \times 10 = 112{,}000$ responses. All collection occurs within the same 15 days to limit temporal drift in the live web index that each provider's search tool consults. Rate limits on the provider side, transient errors, and timeout retries are handled via exponential backoff with a maximum of three retries per call. Calls that fail all three retries are logged with an empty response body and excluded from downstream analysis.

Citation extraction and normalization. The five providers return citations in five different API formats, summarized in Table 6. We parse each format and normalize all citations to a common schema $\langle \texttt{cited_sentence}, \texttt{source_url} \rangle$, where `cited_sentence` is the portion of the answer that relies on the source and `source_url` is its URL. For providers that use inline markers (OpenAI, xAI, Perplexity), we extract a context window of two sentences around each marker. We chose this window based on 389,322 citation spans explicitly delimited by Anthropic and Google,

Table 4: Per-site statistics for the 28 selected Stack Exchange communities. **Category** follows the official Stack Exchange Network six-category taxonomy; **Audience** is taken verbatim from the Sites API `audience` field (fetched $2026{-}04{-}25$); **Pool** reports queries remaining after Step 5 of the quality filter; **Sample** is uniform 400 across all sites; **Year** is the mean posting year within the pool, rounded to the nearest integer.

Category	Site	Audience	Pool	Sample	Year
Technology	Stack Overflow	professional and enthusiast programmers	385,770	400	2020
Technology	Information Security	information security professionals	3,037	400	2020
Technology	Software Engineering	professionals, academics, students in SDLC	2,231	400	2020
Technology	Database Administrators	database professionals	2,692	400	2020
Technology	Bitcoin	Bitcoin users, developers, enthusiasts	816	400	2020
Science	Physics	active researchers, academics, students of physics	11,326	400	2021
Science	Chemistry	scientists, academics, teachers, students	2,981	400	2020
Science	Earth Science	geology, meteorology, oceanography, env. sciences	960	400	2020
Science	Economics	those who study/teach/research economics	622	400	2020
Science	Cross Validated	statistics, ML, data analysis, mining, viz	9,952	400	2020
Science	Artificial Intelligence	people interested in conceptual questions on AI	1,224	400	2020
Science	Medical Sciences	medical and allied health professionals	422	400	2020
Life & Arts	Seasoned Advice	professional and amateur chefs	1,966	400	2021
Life & Arts	Home Improvement	contractors and serious DIYers	3,358	400	2021
Life & Arts	Personal Finance & Money	people who want to be financially literate	2,739	400	2020
Life & Arts	Parenting	parents, grandparents, nannies, others	554	400	2020
Life & Arts	Academia	academics and those in higher education	7,515	400	2021
Life & Arts	Pets	pet owners, vets, breeders, trainers	481	400	2019
Life & Arts	Law	legal professionals, students, others	3,667	400	2022
Culture & Rec.	Bicycles	people who build/repair/ride bicycles	2,556	400	2021
Culture & Rec.	Skeptics	scientific skepticism	2,245	400	2020
Culture & Rec.	Motor Vehicle Maint.	mechanics and DIY car owners	473	400	2021
Culture & Rec.	Travel	road warriors and seasoned travelers	5,849	400	2021
Culture & Rec.	The Great Outdoors	outdoor enthusiasts and learners	1,171	400	2020
Culture & Rec.	Politics	people interested in governments and policies	6,308	400	2021
Professional	The Workplace	members of the workforce	5,148	400	2020
Professional	Aviation	aircraft pilots, mechanics, enthusiasts	4,964	400	2020
Business	Quantitative Finance	finance professionals and academics	849	400	2020
Total			**471,876**	**11,200**	**2020**

Table 5: API configuration for the ten search-augmented LLMs. Search tool is the web-search tool exposed through the provider's API. Perplexity Sonar models perform retrieval natively and require no tool specification. Temperature is N/A for the GPT-5 family, where the parameter is not exposed in the Responses API.

Provider	Model	Search tool	Temp.	Max tokens	Window
OpenAI	gpt-5-2025-08-07	web_search	N/A	8192	03/26–04/09
OpenAI	gpt-5-mini-2025-08-07	web_search	N/A	8192	03/26–04/09
Anthropic	claude-sonnet-4-6	web_search_20250305	0	8192	03/26–04/09
Anthropic	claude-haiku-4-5-20251001	web_search_20250305	0	8192	03/26–04/09
Google	gemini-3-flash	google_search	0	8192	03/26–04/09
Google	gemini-3.1-pro	google_search	0	8192	03/26–04/09
xAI	grok-4-1-fast-non-reasoning	web_search	0	8192	03/26–04/09
xAI	grok-4-1-fast-reasoning	web_search	0	8192	03/26–04/09
Perplexity	sonar	built-in	0	8192	03/26–04/09
Perplexity	sonar-reasoning-pro	built-in	0	8192	03/26–04/09

which average 1.46 sentences, with 91.4% containing at most two sentences. After normalization we obtain $1{,}271{,}046$ citation pairs across the full $112{,}000$ responses. $8{,}069$ responses contain no citations and are excluded from evaluation on individual dimensions but retained in the public release.

Extraction reliability. For providers that return citations as structured blocks (Anthropic, Google), extraction is deterministic because the API returns JSON fields (`cited_text`, `source_url`) that require no heuristic parsing. For providers that use inline markers (OpenAI, xAI, Perplexity), the context window of two sentences captures more than the minimum grounding context, so fidelity judgments (Answer-Source Fidelity, §C.3) remain conservative. A wider window may include text that is not grounded but will not miss the grounded claim. Residual parsing failures such as truncated sentences or malformed markers produce defective `cited_sentence` entries. These entries are caught by the evaluability filters in §B.3, which remove code or table content, overly short extractions, and entries under five words before evaluation on individual dimensions.

Table 6: Citation extraction and normalization by provider.

Provider	Type	Citation unit	URL source	Sentences
OpenAI	Marker	`annotation indices`	`annotations[].url`	Last 2 before marker
xAI	Marker	`annotation indices`	`annotations[].url`	Last 2 before marker
Perplexity	Marker	`[N]` marker	`citations[]` array	Last 2 before marker
Anthropic	Block	`cited_text`	`citations[].url`	Last 2 from block
Google	Block	`grounding_supports[]`	`grounding_chunks[].web.uri`	Last 2 from block

B.3 Crawling and Verifying Source Content

Pipeline. Each unique URL extracted from citations is processed through a pipeline of four stages. First, we resolve redirects (e.g., Gemini proxy URLs) and strip tracking parameters to obtain the canonical URL. Second, we render the page using a headless browser (Playwright Chromium), falling back to an async HTTP client when rendering fails. Third, we convert the retrieved HTML to plain text, applying `trafilatura` (Barbaresi, 2021), `readability-lxml`, and raw `innerText` as successive fallbacks to maximize extraction coverage. Finally, we detect error pages by checking whether the extracted text is too short or matches known patterns for bot blocking. All crawling respects `robots.txt` and enforces rate limiting per domain(Koster et al., 2022). Table 7 lists the specific thresholds and timeout values used at each stage.

Table 7: Crawling pipeline parameters.

Stage	Parameter	Value
URL resolution	Timeout	10 s
Page collection	Timeout per URL	15 s
	Global concurrency	5
	Content cap	50,000 chars
Error detection	Fail if length below	50 chars
	Bot blocking patterns	16 signatures
Rate limiting	Delay per domain	2 s

Table 8: Crawl pipeline outcome by unique URL and citation pair. Inline bars visualize Success and Failed shares for each unit.

Status	URLs	%	Share	Citations	%	Share
Success	231,105	**58.3**		802,945	**63.2**	
Failed	165,565	**41.7**		468,101	**36.8**	
Total	**396,670**	**100.0**		**1,271,046**	**100.0**	

Table 9: Failure category breakdown for crawl pipeline. Highlighted cells mark the two largest categories (JS rendering / bot blocking and unsupported file formats), which together account for 74.7% of all crawl failures.

Failure category	URLs	%	Share	Citations	%	Share
JS rendering / bot blocking	80,517	48.6		248,430	53.1	
File format (PDF / Office)	43,165	26.1		119,592	25.6	
Empty response	29,290	17.7		68,787	14.7	
Server error / access denied	6,662	4.0		16,837	3.6	
Timeout	3,125	1.9		7,316	1.6	
Other (menu / login pages)	2,018	1.2		5,130	1.1	
DNS resolution / domain expiry	788	0.5		2,009	0.4	
Total	**165,565**	**100.0**		**468,101**	**100.0**	

Table 10: Crawl failure rates by host tier (top) and query category (bottom). Forum/Q&A and Social/Blog tiers fail at substantially higher rates due to bot blocking, leading to under-representation of community-based and personal-author sources in the evaluable pool.

Host tier	Total	Failed	%	Share
Forum / Q&A	72,435	49,152	67.9	
Social / Blog	28,146	15,490	55.0	
Academic	52,657	19,489	37.0	
Gov / Org	28,704	9,545	33.3	
Commercial / Other	246,331	70,487	28.6	
News	7,138	1,402	19.6	
Total	**435,411**	**165,565**	**38.0**	

(a) By host tier (URL level)

Query category	Total	Failed	%	Share
Business	43,975	21,002	47.8	
Culture & Recreation	275,576	111,815	40.6	
Professional	78,567	31,896	40.6	
Science	320,204	127,185	39.7	
Life & Arts	349,622	124,075	35.5	
Technology	203,102	52,128	25.7	
Total	**1,271,046**	**468,101**	**36.8**	

(b) By query category (citation level)

Crawl results. The pipeline successfully retrieves 231,105 of 396,670 unique URLs (58.3%), covering 63.2% of all citation pairs (Table 8). The most common reasons for failure are JavaScript rendering or bot blocking and documents in formats other than HTML such as PDF (Table 9). Failures are not uniform across host types. Forum and Q&A hosts fail most often, primarily due to Cloudflare bot blocking, while news domains have the highest success rate (Table 10). As a result, community and personal sources are underrepresented in the evaluable pool. For example, the share of forum and Q&A sources drops from 16.6% of all sources to 8.6% after crawling. We therefore treat our SFR estimate as a conservative lower bound, as discussed further in Appendix D.6 and Section 4.

Table 11: Evaluability filtering stages applied to crawled citations.

Filter	Removed	%	Share	Remaining
–	–	–		802,945
Code or table content	28,530	3.55		774,415
Judge unevaluable	6,878	0.86		767,537
Too short ($<$ 20 chars)	4,774	0.62		762,763
Under 5 words	1,268	0.17		761,495
Total removed	**41,450**	**5.16**		**761,495**

Evaluability filtering. Among the 802,945 successfully crawled citations, we apply four filters to determine the final evaluable pool. We report each filter's effect separately rather than as a single aggregate (Table 11). The filters remove sentences that consist solely of code or table fragments, instances without an ASF label (Judge unevaluable), and sentences that are too short in character length or word count, accounting for 5.16% of the crawled pool. The final evaluable set contains

Table 12: Per-model evaluability statistics, sorted by Eval %. **Eval %** reports each model's evaluable share among its crawled citations; values range from 85.2% (gemini-3-flash) to 99.3% (claude-sonnet). Differences in `cited_code_table` flag rates account for most of the spread. The top performer (claude-sonnet) is highlighted; the bottom row (gemini-3-flash) is shown in bold. *Inline bars use an anchored scale (length = Eval% − 80) rather than the 1% = 1.2pt scale of preceding tables, because the narrow 85–99% range would otherwise make all bars visually identical.*

Model	Evaluable	code_table	judge_un	too_short	lt5_words	Eval %	Share
claude-sonnet-4-6	129,936	223	399	263	58	99.3	
claude-haiku-4-5	37,142	4	147	176	44	99.0	
gpt-5-mini	54,928	50	373	104	36	99.0	
sonar-reasoning-pro	69,411	208	319	176	45	98.9	
gpt-5	38,322	33	294	146	53	98.7	
gemini-3.1-pro	77,999	191	559	832	86	97.9	
grok-4-1-fast-reasoning	92,355	5,494	1,080	601	170	92.6	
grok-4-1-fast-non-reasoning	90,916	5,843	1,181	623	229	92.0	
sonar	122,157	12,042	942	434	106	90.0	
gemini-3-flash	48,329	4,452	2,073	1,420	442	85.2	
Total	**761,495**	**28,540**	**7,367**	**4,775**	**1,269**	**94.8**	

Table 13: Sensitivity of failure rates to reinstating 2,633 excluded instances retaining partial prose. Since these instances lack labels, they are treated as either all non-failures or all failures.

Metric	Reported (%)	All non-failures (%)	All failures (%)
Overall SFR	27.12	27.03	27.37
Overall FFR	30.60	30.49	30.84
Code/Data FFR	31.72	31.30	32.65
Technical FFR	32.23	32.05	32.60

761,495 citation triples across the ten models. Evaluable rates per model range from 85.2% (Gemini 3 Flash) to 99.3% (Claude Sonnet), with variation largely due to differences in how often each model produces code in its citations (Table 12). All metrics in Appendix D are computed on this filtered pool.

We exclude code/table fragments and very short text because they may lack sufficient prose for fidelity assessment. We audited 6,878 instances without an ASF label. Of these, 90.9% contained headings, table fragments, or bare numerals without an adjudicable proposition, and 6.7% had an intact citing sentence but an access notice or copyright banner as source content. Another 164 instances contained an adjudicable proposition, equivalent to approximately 0.02% of the evaluable pool.

Separately, we assessed the effect of reinstating 2,633 instances retaining partial prose from the 28,530 excluded by the code/table filter. Since these instances lacked labels, we treated them as all non-failures or all failures (Table 13). SFR ranges from 27.03% to 27.37%, and FFR from 30.49% to 30.84%, indicating that reinstating these instances has a limited effect on aggregate failure rates.

C Three-Dimension Evaluation Framework

This appendix expands the evaluation framework summarized in Section 3. Following the structure of the main text, we describe Intent–Purpose Alignment in §C.1, Source Suitability in §C.2, and Answer–Source Fidelity in §C.3. For each dimension we present the taxonomy, the scoring rubric, the aggregation method at the response level, and the supporting validation evidence. Integration across dimensions and robustness checks are covered in §C.4. Each dimension is built from one or

two underlying classification tasks: Query Intent (QI) and Source Purpose (SP) for alignment, Source Domain (SD) and Source Type (ST) for suitability, and Answer–Source Fidelity (ASF) for fidelity.

Judge configuration. All five classification tasks use a single LLM judge (gpt-4o-mini-2024-07-18), called with temperature 0, strict JSON-mode structured output, and a $4{,}096$-token output limit. The full prompt for each task appears in Figures 7 through 10.

C.1 Alignment Between Query Intent and Source Purpose

Query Intent (QI) taxonomy. Query Intent classifies each query by its *primary completion condition* (Taylor, 1968; Broder, 2002; Ingwersen & Järvelin, 2005), defined as what the user must obtain to consider the query resolved. Drawing on prior question-type and search-intent taxonomies (Bolotova et al., 2022; Rose & Levinson, 2004; Jansen et al., 2008), we define five labels (Table 14): QI1 Factoid (single verifiable datum), QI2 Explanation (causal mechanism or principle), QI3 Instruction (procedural steps or troubleshooting), QI4 Comparison (evaluation of alternatives), and QI5 Opinion (subjective or value-laden judgment).

Table 14: Query Intent (QI) taxonomy: five labels classifying each query by its primary completion condition.

Label	Name	Definition
QI1	Factoid	A query seeking a specific, verifiable datum that is fully resolved by a single lookup with a context-independent answer.
QI2	Explanation	A query seeking the causal mechanism or underlying principle behind an observed phenomenon.
QI3	Instruction	A query seeking procedural steps or methods to perform a task or resolve a discrepancy, including troubleshooting.
QI4	Comparison	A query seeking evaluation of alternatives against explicit, intersubjectively shareable criteria to support a decision or recommendation.
QI5	Opinion	A query seeking a subjective, ethical, or value-laden judgment on a socially contested question.

Source Purpose (SP) taxonomy. Source Purpose classifies each cited source by its *communicative function* (Biber et al., 2015; Sharoff, 2018; Swales, 1990), defined as the goal that the source's author pursues regardless of factual content. We define six labels (Table 15): SP1 To Promote (commercial advocacy), SP2 To Inform (neutral knowledge transmission), SP3 To Instruct (procedural guidance), SP4 To Report (event reporting), SP5 To Discuss (community deliberation), and SP6 To Opine (individual perspective). This distinction matters because the purpose of a source shapes its structural incentive: a promotional source may present accurate facts yet frame them to favor a product, creating a mismatch when paired with informational queries.

System prompt

```
You are a search-augmented LLM.
Use the search tool to retrieve relevant external knowledge
and provide an accurate, comprehensive answer to the following
question.
```

User prompt

```
{query}
```

Figure 6: Prompt for response generation

Table 15: Source Purpose (SP) taxonomy: six labels classifying each source by its communicative function.

Label	Name	Definition
SP1	To Promote	A source that advocates for a specific product, service, or commercial entity to drive purchase or favorable perception.
SP2	To Inform	A source that transmits factual or conceptual knowledge with no commercial incentive. Encyclopedic entries, official documentation.
SP3	To Instruct	A source that provides procedural guidance: tutorials, manuals, how-to guides.
SP4	To Report	A source that reports an event, observation, or finding: news articles, study reports.
SP5	To Discuss	A source hosting community discussion or deliberation: forum threads, Q&A sites.
SP6	To Opine	A source presenting an individual perspective or opinion: op-eds, personal blogs, commentary.

Table 16: Intent–Purpose Alignment (IPA) Matrix. Each cell scores alignment between Query Intent (rows) and Source Purpose (columns) from 1 (structural conflict) to 5 (direct match). Red cells (≤ 2) define alignment failures contributing to the AFR metric. Matrix values are validated by 10 IR/LIS researchers ($\text{ICC}(2, k) = 0.916$; Table 25).

	SP1 To Promote	**SP2** To Inform	**SP3** To Instruct	**SP4** To Report	**SP5** To Discuss	**SP6** To Opine
QI1 Factoid	1	5	3	4	2	1
QI2 Explanation	1	5	4	3	3	2
QI3 Instruction	2	3	5	2	4	2
QI4 Comparison	3	4	2	3	3	3
QI5 Opinion	2	3	2	3	4	5

IPA matrix design. The Intent–Purpose Alignment(IPA) Matrix is a 5×6 matrix that scores each (QI, SP) combination on a 1–5 scale (Table 16). For a citation triple $c = (q, a_c, s_c)$ classified as $\mathrm{QI}(q)$ and $\mathrm{SP}(s_c)$, the alignment score is

$$\mathrm{IPA}(c) = \mathbf{M}_{IP}[\mathrm{QI}(q),\, \mathrm{SP}(s_c)] \in \{1, 2, 3, 4, 5\}. \tag{1}$$

Scores of 3–5 indicate functional alignment, where the source's communicative function contributes to the query's completion condition (5 = direct match, 3 = partial relevance). Scores of 1–2 indicate structural misalignment, where the source's incentive conflicts with the query's information need (1 = structural conflict, 2 = weak fit). We define citations with $\mathrm{IPA}(c) \leq 2$ as alignment failures. This threshold corresponds to the boundary between the two regimes in the rubric, and we verify in Appendix D.6 that shifting the threshold by ± 1 preserves relative model rankings. Per-cell justifications appear in Table17–18.

Response-level aggregation. For a response r with citation set C_r, we define two metrics:

$$\text{R-IPA}(r) = \frac{1}{|C_r|} \sum_{c \in C_r} \mathrm{IPA}(c), \qquad \mathrm{AFR}(r) = \frac{|\{c \in C_r : \mathrm{IPA}(c) \leq 2\}|}{|C_r|}, \tag{2}$$

$$\text{R-AFR}(r) = \mathbb{1}\big[\exists\, c \in C_r : \mathrm{IPA}(c) \leq 2\big]. \tag{3}$$

R-IPA is the mean alignment score per response (higher is better), AFR (Alignment Failure Rate) is the share of citations in the conflict region (lower is better), and R-AFR (Response-level Alignment Failure Rate) is a binary indicator of whether the user encounters at least one misaligned citation in the response. Responses with zero citations are excluded from aggregation but retained in the public release.

Validation. A panel of 10 IR/LIS researchers independently rated all 30 IPA Matrix cells on a 1–5 scale. Inter-rater reliability is excellent ($\mathrm{ICC}(2, k) = 0.916$ (Koo & Li, 2016), 95% CI $= [0.86, 0.95]$), and the correlation between expert consensus and design values is strong ($r = 0.871$, $\mathrm{CCC} = 0.821$). No cell deviates from expert consensus by 2 or more points, and the four cells with deviations of 1.0–1.2 points are all in the conservative direction. Full validation statistics appear in Table 25. On the QI ($\kappa = 0.862$) and SP ($\kappa = 0.862$) label-assignment tasks, the LLM judge reaches substantial agreement against three human annotators (Table 27).

C.2 Source Suitability Across Domains and Types

Source Domain (SD) taxonomy. SD classifies each cited source by the substantive domain of its content into ten labels (IAB Tech Lab, 2022) (Table 19): SD1 Medical (diseases, treatments, mental health), SD2 Legal (law, regulation), SD3 Finance (banking, taxation), SD4 Education (curriculum, university, scholarship), SD5 Science (natural science, research), SD6 Code/Data (software, datasets), SD7 Technical (engineering, specifications), SD8 Social/Professional (workplace, career), SD9 Shopping/Travel (commerce, lifestyle purchases), and SD10 Everyday (recipes, hobbies, routine information). Labels SD1–SD3 correspond to the YMYL domains identified by Google's Search Quality Rater Guidelines (Google, 2025); SD4–SD7 capture knowledge-intensive domains; SD8–SD10 capture lifestyle and routine information.

Source Type (ST) taxonomy. ST classifies each source by its publication infrastructure and editorial accountability (Rieh & Danielson, 2007; Sun et al., 2019) into six labels (Table 20): ST1 Official (government, regulatory agencies, nonprofits), ST2 Research (peer-reviewed academic publication), ST3 News (mainstream journalism), ST4 Wiki/Forum (community-edited or community-discussion content), ST5 Blog/Social (individual-author or social-media content), and ST6 Company (corporate or commercial content). This is distinct from SP, which captures communicative intent rather than institutional structure.

SS matrix design. The Source Suitability (SS) Matrix is a 10×6 matrix scoring each (SD, ST) combination on a 1–5 scale (Table 21). For citation triple $c = (q, a_c, s_c)$ with source s_c, the suitability score is

$$\mathrm{SS}(c) = \mathbf{M}_{SS}[\mathrm{SD}(s_c),\, \mathrm{ST}(s_c)] \in \{1, 2, 3, 4, 5\}. \tag{4}$$

Table 17: Full cell-level justifications for the IPA matrix (Part 1: QI1–QI3). Each row explains why the (QI, SP) pair receives its assigned score on the 1–5 scale.

Cell	Score	Rationale
QI1 Factoid		
QI1×SP1	1	Factoid queries require a context-independent verifiable fact; a promotional source's incentive to selectively present favorable information introduces adverse selection even when the fact itself is accurate.
QI1×SP2	5	Neutral knowledge-transmission sources (encyclopedias, documentation) directly resolve factoid queries without incentive distortion.
QI1×SP3	3	Instructional sources may contain the target fact within procedural context, but it is embedded rather than foregrounded; partial relevance.
QI1×SP4	4	News reports frequently contain verifiable facts with editorial accountability, providing substantial support for factoid resolution.
QI1×SP5	2	Community discussions may surface correct facts but lack editorial accountability; the user cannot distinguish verified from anecdotal claims.
QI1×SP6	1	Opinion sources foreground subjective judgment; citing an opinion piece to resolve a factoid query structurally misrepresents the evidentiary basis.
QI2 Explanation		
QI2×SP1	1	Explanation queries seek causal understanding; promotional sources selectively frame mechanisms to favor their product, distorting the explanatory account.
QI2×SP2	5	Informational sources are designed to transmit causal mechanisms and principles, directly matching the explanation completion condition.
QI2×SP3	4	Instructional sources often embed explanatory content (why a step works), providing substantial though not primary explanatory value.
QI2×SP4	3	News reports may describe causes of events but typically lack the depth required for mechanistic understanding; partial relevance.
QI2×SP5	3	Community discussions may contain expert-level explanations but are mixed with speculative or incomplete accounts; partial relevance.
QI2×SP6	2	Opinion pieces may offer interpretive frameworks but foreground the author's perspective over neutral mechanism; weak fit for explanation.
QI3 Instruction		
QI3×SP1	2	Promotional sources may include product-specific procedures, but the commercial incentive to steer users toward a particular solution weakens procedural neutrality.
QI3×SP2	3	Informational sources provide background knowledge but rarely offer executable step-by-step procedures; the user must infer the procedure from conceptual content.
QI3×SP3	5	Instructional sources (tutorials, how-to guides) directly match the procedural completion condition of instruction queries.
QI3×SP4	2	News reports describe events rather than actionable procedures; citing a report to resolve an instructional query mismatches the source function.
QI3×SP5	4	Community discussions (e.g., Stack Overflow answers) frequently provide tested, peer-reviewed procedures; substantial procedural support.
QI3×SP6	2	Opinion sources express preferences rather than executable procedures; weak fit for resolving a procedural query.

Table 18: Full cell-level justifications for the IPA matrix (Part 2: QI4–QI5). Each row explains why the (QI, SP) pair receives its assigned score on the 1–5 scale.

Cell	Score	Rationale
QI4 Comparison		
QI4×SP1	3	Promotional sources may contain comparative claims, but the commercial incentive biases the comparison; partially relevant if the user seeks feature lists.
QI4×SP2	4	Informational sources provide neutral feature descriptions that support structured comparison, though they may not directly rank alternatives.
QI4×SP3	2	Instructional sources focus on how to use a single option rather than comparing alternatives; weak fit for evaluative comparison.
QI4×SP4	3	News reports may cover product launches or policy alternatives, offering partial comparative context without structured evaluation.
QI4×SP5	3	Community discussions may contain experiential comparisons, but the lack of systematic criteria limits their evaluative reliability; partial relevance.
QI4×SP6	3	Opinion sources may express ranked preferences, but subjectivity and lack of sharable criteria place them at partial relevance for structured comparison.
QI5 Opinion		
QI5×SP1	2	Promotional sources may contain implicit value positions, but their commercial incentive makes them unreliable representatives of genuine opinion diversity.
QI5×SP2	3	Informational sources may present multiple perspectives neutrally but do not themselves take a position; partial relevance for opinion-seeking queries.
QI5×SP3	2	Instructional sources address procedural tasks, not value judgments; citing a how-to guide for an opinion query mismatches the source function.
QI5×SP4	3	News reports may frame contested issues and present stakeholder positions, offering partial context for forming opinions.
QI5×SP5	4	Community discussions surface diverse perspectives from experienced participants, providing substantial support for opinion formation.
QI5×SP6	5	Opinion sources directly match the completion condition of opinion queries: the user seeks a subjective, value-laden judgment, which is precisely what opine sources provide.

Table 19: Source Domain taxonomy: ten labels classifying the topical domain of each cited source.

Label	Domain	Definition
SD1	Medical/Health	Diseases, treatments, medications, mental health, nutrition
SD2	Legal	Laws, regulations, court decisions, legal rights, compliance
SD3	Finance	Personal finance, investing, economics, taxation, banking
SD4	Education	Education, curriculum, university, degree, scholarship
SD5	Science	Natural sciences, mathematics, physics, chemistry, biology
SD6	Code/Data	Programming, software, data analysis, machine learning, AI
SD7	Technical	IT systems, infrastructure, cloud services, mechanics
SD8	Social/Professional	Society, relationships, workplace, career, parenting
SD9	Shopping/Travel	Shopping, product reviews, travel, accommodation
SD10	Everyday	Daily life, DIY, hobby, lifestyle, sports, pets, cooking

Table 20: Source Type taxonomy: six labels classifying the institutional type of each cited source.

Label	Name	Definition
ST1	Official Institution	Government, regulatory agencies, nonprofits, academic institutions
ST2	Paper/Research	Peer-reviewed academic paper with author, abstract, references
ST3	News/Magazine	News article with byline and publication date
ST4	Wiki/Forum	Community-created content: wikis, Q&A, forums
ST5	Blog/Social	Individual-authored content: blogs, social media, personal pages
ST6	Private Company	Company-published content: product pages, docs, corporate blogs

Cell values follow the CRAAP test (Blakeslee, 2004) for source evaluation and the YMYL classification (Google, 2025), governed by two principles. First, in YMYL domains (SD1–SD3), only Official and Research sources receive 5, while Wiki/Forum and Blog/Social receive 1 in Medical and Legal domains due to the risk of unverified information in high-stakes contexts. Second, in Code/Data (SD6), Wiki/Forum sources (ST4) receive 5 because developer communities such as Stack Overflow hold de facto authority in software engineering; this reflects the empirical authority structure of the domain rather than a deviation from the CRAAP framework. In non-YMYL domains (SD9 Shopping/Travel, SD10 Everyday), the minimum cell value is 3 rather than 1, reflecting the lower stakes of misalignment in routine contexts. We define citations with $\mathrm{SS}(c) \leq 2$ as suitability failures, and verify in Appendix D.6 that shifting the threshold by ± 1 preserves relative model rankings. Per-cell justifications appear in Tables 22–24.

Table 21: Source Suitability Matrix (SS Matrix). Scores range from 1 (structural mismatch) to 5 (strong match). Red cells (≤ 2) indicate suitability failures.

	ST1 Official	**ST2** Research	**ST3** News	**ST4** Wiki/Forum	**ST5** Blog/Social	**ST6** Company
SD1 Medical	5	5	3	1	1	2
SD2 Legal	5	5	3	1	1	2
SD3 Finance	5	5	3	2	2	3
SD4 Education	5	5	3	3	2	3
SD5 Science	5	5	3	3	2	2
SD6 Code/Data	4	4	3	5	2	3
SD7 Technical	5	4	3	3	2	3
SD8 Social	4	4	4	3	3	3
SD9 Shopping	4	3	4	4	3	3
SD10 Everyday	4	3	4	4	4	3

Response-level aggregation. For a response r with citation set C_r, we define two metrics:

$$\text{R-SS}(r) = \frac{1}{|C_r|} \sum_{c \in C_r} \mathrm{SS}(c), \qquad \mathrm{SFR}(r) = \frac{|\{c \in C_r : \mathrm{SS}(c) \leq 2\}|}{|C_r|}. \tag{5}$$

$$\text{R-SFR}(r) = \mathbb{1}\big[\exists\, c \in C_r : \mathrm{SS}(c) \leq 2\big]. \tag{6}$$

R-SS is the mean suitability score per response (higher is better), SFR (Suitability Failure Rate) is the share of citations from unsuitable source types for the relevant domain (lower is better), and R-SFR (Response-level Suitability Failure Rate) is a binary indicator of whether the user encounters at least one unsuitable source in the response. Responses with zero citations are excluded from aggregation but retained in the public release.

Table 22: Full cell-level justifications for the SS matrix (Part 1: YMYL domains SD1–SD3). Each row explains why the (SD, ST) pair receives its assigned score.

Cell	Score	Rationale
SD1 Medical (YMYL)		
SD1×ST1	5	Government health agencies (CDC, WHO, NHS) undergo peer review and regulatory oversight; highest reliability for medical claims.
SD1×ST2	5	Peer-reviewed medical research is the gold standard for clinical evidence; directly appropriate for medical queries.
SD1×ST3	3	Mainstream health journalism may simplify or sensationalize findings but operates under editorial accountability; moderate reliability.
SD1×ST4	1	Wiki/forum content in medical domains carries high misinformation risk; anonymous contributors lack clinical accountability.
SD1×ST5	1	Individual health blogs routinely propagate unverified claims with no editorial or clinical oversight.
SD1×ST6	2	Pharmaceutical and health-product companies have regulatory disclosure obligations but also commercial incentives that bias health claims.
SD2 Legal (YMYL)		
SD2×ST1	5	Official legal sources (statutes, court opinions, regulatory guidance) are authoritative by definition in legal domains.
SD2×ST2	5	Legal scholarship in peer-reviewed journals provides reliable doctrinal analysis and case interpretation.
SD2×ST3	3	Legal journalism provides accessible summaries but may lack jurisdictional precision; moderate reliability.
SD2×ST4	1	Forum-based legal advice from non-lawyers carries high risk of jurisdictional error and misapplied precedent.
SD2×ST5	1	Personal legal blogs may offer opinions without bar admission or malpractice accountability; structurally inappropriate for legal guidance.
SD2×ST6	2	Corporate legal content has professional accountability but also client-acquisition incentives.
SD3 Finance (YMYL)		
SD3×ST1	5	Financial regulators (SEC, central banks) provide authoritative data and guidance with statutory accountability.
SD3×ST2	5	Peer-reviewed finance research offers rigorous, methodologically transparent analysis of financial phenomena.
SD3×ST3	3	Financial journalism provides timely market coverage under editorial standards but may amplify short-term sentiment.
SD3×ST4	2	Finance forums mix informed analysis with speculative advice; mostly inappropriate for financial decisions.
SD3×ST5	2	Personal finance blogs vary widely in quality; absence of fiduciary duty makes them mostly inappropriate for financial guidance.
SD3×ST6	3	Financial institutions' content is subject to regulatory disclosure (10-K, prospectuses) but carries sales incentives; moderate reliability.

Table 23: Full cell-level justifications for the SS matrix (Part 2: knowledge-intensive domains SD4–SD7). Each row explains why the (SD, ST) pair receives its assigned score.

Cell	Score	Rationale
SD4 Education		
SD4×ST1	5	Official educational bodies (ministries, accreditation agencies) set authoritative curricular and policy standards.
SD4×ST2	5	Peer-reviewed educational research provides evidence-based pedagogical guidance.
SD4×ST3	3	Education journalism covers policy and trends but may lack pedagogical depth; moderate relevance.
SD4×ST4	3	Educational wikis and forums provide useful crowd-sourced study resources but lack systematic quality control.
SD4×ST5	2	Individual education blogs may share teaching tips but lack institutional backing or peer review; mostly inappropriate.
SD4×ST6	3	EdTech companies provide product-linked educational content with commercial incentives but often useful pedagogical material.
SD5 Science		
SD5×ST1	5	Government science agencies (NASA, NOAA, NIH) provide authoritative scientific data and reports.
SD5×ST2	5	Peer-reviewed scientific publications are the primary knowledge-creation mechanism in science.
SD5×ST3	3	Science journalism translates findings for lay audiences but may oversimplify or misrepresent effect sizes.
SD5×ST4	3	Science wikis and forums (e.g., Physics Stack Exchange) contain expert-level content but lack formal peer review.
SD5×ST5	2	Science blogs may distort findings through selective reporting or lack of methodological training; mostly inappropriate.
SD5×ST6	2	Corporate science content carries commercial incentives that may bias interpretation.
SD6 Code/Data		
SD6×ST1	4	Official documentation (language specs, RFC standards) is authoritative but may lag behind practical usage patterns.
SD6×ST2	4	Academic CS publications provide algorithmic foundations but may not address practical implementation details.
SD6×ST3	3	Tech journalism covers releases and trends but rarely provides actionable code-level guidance.
SD6×ST4	5	Developer communities (Stack Overflow, GitHub Issues) are the de facto authority for code solutions; peer-tested and version-specific.
SD6×ST5	2	Individual coding blogs vary widely; outdated or untested code snippets may introduce bugs; mostly inappropriate as primary sources.
SD6×ST6	3	Company developer docs and SDKs are authoritative for their own products but may not generalize; moderate reliability.
SD7 Technical		
SD7×ST1	5	Official standards bodies (ISO, IEEE, building codes) provide authoritative technical specifications.
SD7×ST2	4	Engineering research publications provide validated methods but may not cover practical implementation constraints.
SD7×ST3	3	Technical journalism provides accessible overviews but lacks specification-level precision.
SD7×ST4	3	Technical forums contain practitioner knowledge but lack formal verification; moderate reliability.
SD7×ST5	2	Individual technical blogs may share useful experience but lack institutional backing or peer review.
SD7×ST6	3	Manufacturer documentation is authoritative for specific products but may omit cross-product comparisons.

Table 24: Full cell-level justifications for the SS matrix (Part 3: lifestyle and routine domains SD8–SD10). Non-YMYL domains use a minimum score of 3, reflecting lower stakes of source-type misalignment.

Cell	Score	Rationale
SD8 Social/Professional		
SD8×ST1	4	Government labor agencies provide authoritative employment data and workplace regulations.
SD8×ST2	4	Organizational-behavior and HR research provides evidence-based workplace guidance.
SD8×ST3	4	Workplace and career journalism provides timely, editorially reviewed advice on professional norms.
SD8×ST4	3	Workplace forums offer experiential advice but norms vary by culture and industry.
SD8×ST5	3	Career blogs share individual experiences that may not generalize but carry lower risk than YMYL domains.
SD8×ST6	3	Corporate HR content is informative but self-promotional.
SD9 Shopping/Travel		
SD9×ST1	4	Government tourism and consumer-protection agencies provide reliable destination and product-safety information.
SD9×ST2	3	Academic tourism or consumer research is reliable but rarely addresses practical purchasing or travel decisions.
SD9×ST3	4	Travel and product journalism provides timely, editorially reviewed recommendations with moderate commercial influence.
SD9×ST4	4	Community review sites and travel forums provide firsthand, peer-validated consumer experiences.
SD9×ST5	3	Individual travel and shopping blogs provide personal experience; commercial sponsorship is common but stakes are low.
SD9×ST6	3	E-commerce and travel-company sites provide product details with obvious sales incentives; useful but biased.
SD10 Everyday		
SD10×ST1	4	Government agencies provide authoritative guidance on everyday matters (food safety, recycling rules).
SD10×ST2	3	Academic research on everyday topics is reliable but often overly technical for casual queries.
SD10×ST3	4	Lifestyle journalism provides accessible, editorially reviewed content on everyday topics.
SD10×ST4	4	Community wikis and forums are well-suited for everyday advice where collective experience is the primary authority.
SD10×ST5	4	Personal lifestyle blogs are a natural fit for casual everyday queries where no formal authority is required.
SD10×ST6	3	Corporate lifestyle content is useful but shaped by commercial incentives.

Validation. A panel of 100 domain experts (10 per domain) independently rated the 6 cells of their assigned domain on a 1–5 scale. SD1–SD7 were rated by domain professionals or researchers, and SD8–SD10 by graduate-level researchers. Inter-rater reliability is excellent across domains (9 of 10 domains with $\mathrm{ICC}(2,k) \geq 0.90$ (Koo & Li, 2016), median $= 0.958$), and the correlation between expert consensus and design values is strong ($r = 0.943$, $\mathrm{CCC} = 0.934$). No cell exhibits Tier 1 disagreement (≥ 2 points from expert consensus). Full per-domain statistics appear in Table 25.

Table 25: Expert validation of IPA and SS matrices. **ICC**$(2,k)$ is two-way mixed-effects intraclass correlation for absolute agreement of average measurements; **Pearson** r is the correlation between expert consensus and design values; **MAD** is mean absolute deviation on the 1-to-5 scale. The IPA aggregate row reports validation by $n = 10$ experts on all 30 cells; SS rows report per-domain validation by $n = 10$ experts on the 6 cells of each domain, and the SS aggregate row pools all 60 cells. The **YMYL** column marks the three high-stakes Source Domains (Medical, Legal, and Finance) with ✓, and their rows are highlighted accordingly.

Matrix	Row	YMYL	N	ICC$(2,k)$	95% CI	Pearson r	MAD
IPA	Aggregate (30 cells)		10	0.916	[0.86, 0.95]	0.871	0.513
SS	SD1 Medical	✓	10	0.980	[0.94, 1.00]	0.990	0.350
	SD2 Legal	✓	10	0.983	[0.95, 1.00]	0.990	0.350
	SD3 Finance	✓	10	0.973	[0.92, 1.00]	0.969	0.267
	SD4 Education		10	0.951	[0.86, 0.99]	0.973	0.250
	SD5 Science		10	0.964	[0.90, 0.99]	0.988	0.250
	SD6 Code/Data		10	0.908	[0.74, 0.98]	0.913	0.350
	SD7 Technical		10	0.931	[0.81, 0.99]	0.937	0.383
	SD8 Social		10	0.972	[0.92, 1.00]	0.842	0.500
	SD9 Shopping		10	0.945	[0.85, 0.99]	0.753	0.300
	SD10 Everyday		10	0.895	[0.71, 0.98]	0.778	0.267
	Aggregate (60 cells)		**100**	**0.962**	**[0.95, 0.97]**	**0.943**	**0.327**

C.3 Fidelity of Answers to Their Sources

Table 26: Answer–Source Fidelity (ASF) taxonomy: five labels adjudicating the fidelity of each cited claim to its source.

Label	Name	Definition
ASF1	Fabricated	The cited claim does not exist in the source content at all.
ASF2	Misattributed	The claim comes from a tangential part or is attributed to a context the source does not address.
ASF3	Contradicted	The source concludes or argues the opposite of what the citing sentence presents.
ASF4	Amplified	The claim exists but is presented with materially greater certainty, scope, or generality.
ASF5	Supported	The claim is present in and consistent with the source content.

Answer–Source Fidelity (ASF) taxonomy. The ASF taxonomy classifies the relationship between a citing sentence and the crawled source content on a 1–5 scale (Table 26): ASF5 Supported (every claim attributable to the source), ASF4 Amplified (claims attributable but a critical qualifier omitted), ASF3 Contradicted (a hedged claim presented as unconditional or reversed), ASF2 Misattributed (the source provides only tangential evidence), and ASF1 Fabricated (no claim supported by the source). The taxonomy extends the binary supported/not-attributable judgment of ALCE (Gao et al., 2023) and AIS (Rashkin et al., 2023; Maynez et al., 2020; Pagnoni et al., 2021; Kryscinski et al.,

Table 27: LLM judge reliability against human consensus across the five classification tasks. Validation uses $n = 200$ stratified samples per task with three human annotators, oversampling rare categories. κ_{cons} is Cohen's κ against majority-vote consensus, κ_{pair} is the mean pairwise κ between the judge and each annotator, and Bal. Acc is balanced accuracy. All five tasks achieve $\kappa_{\text{cons}} \geq 0.78$.

Task	N	Raw Agr	κ_{cons}	95% CI	κ_{pair}	Bal. Acc
QI	200	0.890	**0.862**	$[0.806, 0.912]$	0.817	0.900
SP	200	0.885	**0.862**	$[0.807, 0.915]$	0.824	0.896
SD	200	0.835	**0.817**	$[0.756, 0.872]$	0.783	0.872
ST	200	0.830	**0.796**	$[0.731, 0.856]$	0.772	0.860
ASF	200	0.905	**0.858**	$[0.790, 0.913]$	0.839	0.918

2020; Wang et al., 2020; Laban et al., 2022); the three intermediate levels (ASF4–ASF2) capture the directional distortions characteristic of Verified Misguidance.

Citing sentence and source content. Each citation triple consists of the query, the citing sentence, and the corresponding cited source content. The extraction procedure for both marker and block providers is described in Appendix B.2. The ASF judge receives the citing sentence and the cited source content, and assigns one of the five verdicts (ASF1–ASF5).

Response-level aggregation. For a response r with citation set C_r, we define two metrics:

$$\text{FFR}(r) = \frac{|\{c \in C_r : \text{ASF}(c) \leq 2\}|}{|C_r|}, \qquad \text{R-FFR}(r) = \mathbb{1}\big[\exists\, c \in C_r : \text{ASF}(c) \leq 2\big]. \tag{7}$$

FFR (Fidelity Failure Rate) is the share of citations that fail (lower is better), and R-FFR (Response-level Fidelity Failure Rate) is a binary indicator of whether the user encounters at least one failed citation in the response. The threshold $\text{ASF}(c) \leq 2$ groups Fabricated (ASF1) and Misattributed (ASF2), both of which share the property that a user examining the source cannot independently verify the cited claim, distinguishing them from Contradicted (ASF3) or Amplified (ASF4) cases where the source contains the claim in some form. This threshold does not imply that Contradicted (ASF3) or Amplified (ASF4) citations are faithful. Appendix D.6 reports the sensitivity of CritVM rankings to including ASF3 in the failure definition. Responses with zero citations are excluded from aggregation but retained in the public release.

Validation. The LLM judge's ASF classification was validated against three human annotators on 200 stratified samples (140 comprising 100 Supported and 40 Fabricated cases, plus 60 samples covering three intermediate subtypes at 20 each) to ensure adequate coverage of rare failure categories. The three annotators were co-investigators or domain-expert collaborators with experience in information retrieval and web-search evaluation. They labeled independently using the same category definitions and fidelity rubric provided to the judge, without seeing the judge's labels or the identity of the response-generating model. Cohen's κ (Cohen, 1960) against majority-vote consensus is 0.858 (95% CI $= [0.790, 0.913]$), indicating excellent agreement, with balanced accuracy of 0.918. Agreement among the three human annotators is Krippendorff's $\alpha = 0.809$ (Krippendorff, 2011) (minimum human–human pairwise $\kappa = 0.793$). This validation measures agreement with project-affiliated annotators applying the specified criteria, rather than agreement across annotators from different backgrounds. Full reliability statistics for all five classification tasks appear in Table 27.

C.4 Cross-Dimension Integration and Robustness

Critical VM definition. We define a citation triple as a Critical VM instance when all three dimensions fail simultaneously:

$$\text{CritVM}(c) = \mathbb{1}\big[\text{IPA}(c) \leq 2 \;\wedge\; \text{ASF}(c) \leq 2 \;\wedge\; \text{SS}(c) \leq 2\big]. \tag{8}$$

Each individual dimension condition corresponds to an established evaluation framework: Intent–Purpose Alignment to query intent taxonomies such as NF-CATS (Bolotova et al., 2022), Source Suitability to source quality frameworks such as CRAAP (Blakeslee, 2004), and Answer–Source

Table 28: Cohen's κ against human majority-vote consensus on 200 samples per classification task, using identical task-specific prompts across judge models.

Task	GPT-4o-mini	Claude Opus 4.7	Qwen3.5-Plus
QI	.862	.704	.623
SP	.862	.675	.650
SD	.817	.871	.854
ST	.796	.712	.852
ASF	.858	.463	.429

Fidelity to citation verification benchmarks such as ALCE (Gao et al., 2023) and CiteME (Press et al., 2024). The simultaneous failure across all three is not captured by any single existing framework, and this composition across dimensions is the central evaluative contribution of CITETRACE.

Comparison with the independence expectation. Assuming that the three dimensions fail independently, the expected joint failure rate is the product of the marginal rates:

$$\mathbb{E}_{\text{indep}}[\text{CritVM}] = \text{AFR} \cdot \text{FFR} \cdot \text{SFR} = 0.051 \cdot 0.306 \cdot 0.271 \approx 0.0042 = 0.42\%. \tag{9}$$

The observed CritVM rate is $3{,}174/761{,}495 \approx 0.42\%$, close to the independent-failure expectation.

Single judge dependence. All five classification tasks use a single LLM judge (GPT-4o-mini). This is consistent with the evaluation methodology of comparable frameworks: ALCE (Gao et al., 2023) validates citation quality against a single NLI model ($\kappa = 0.525$), RAGAS (Es et al., 2024) prompts a single LLM, and ARES (Saad-Falcon et al., 2024) trains a single tailored judge per dimension. Our judge achieves Cohen's κ of 0.796–0.862 against human majority-vote consensus (Table 27). The stratified validation protocol (§C.3) with three independent human annotators (Krippendorff's $\alpha \geq 0.809$) supports the reliability of the judge's classifications. Nonetheless, absolute metric values may shift under judge replacement; we discuss this dependence in Appendix A and encourage replication with alternative judge families.

Comparison across judge models. GPT-4o-mini achieved the highest agreement with human consensus on three of the five classification tasks, including ASF. We compared GPT-4o-mini, Claude Opus 4.7, and Qwen3.5-Plus on the same 200 human-validated samples per task using identical task-specific prompts without model-specific tuning. As shown in Table 28, GPT-4o-mini led on QI, SP, and ASF, while Claude and Qwen led on SD and ST, respectively. These results support our choice of GPT-4o-mini for the evaluation pipeline based on human agreement.

Assessment of within-provider judging bias. To examine whether the OpenAI judge scored responses from its own provider more leniently than human annotators, we focused on ASF, the only classification task that directly evaluates model-generated text. For each of the 200 ASF validation samples, we calculated the judge's score minus the human majority-vote score, with positive values indicating more lenient scoring. Among the 34 citation instances from OpenAI models, none received a higher score from the judge than from the human consensus, and the mean deviation was -0.18.

Threshold sensitivity protocol. The failure threshold ≤ 2 for all three dimensions is anchored in the conflict and failure regime built into the rubrics (§C.1, §C.2, §C.3). To quantify sensitivity to this choice, we evaluate the baseline and six ± 1 perturbation variants, using thresholds $\{< 2, \leq 2, \leq 3\}$ for each dimension independently. Model-level CritVM rankings preserve Kendall $\tau \geq 0.82$ in five of the seven settings, with the two exceptions involving the < 2 thresholds for Source Suitability and Intent–Purpose Alignment ($\tau = 0.733$ and 0.644, respectively). Full ranking-sensitivity results appear in Appendix D.6.

System prompt

```
You are a query intent classifier.  Given a user query, you
assign exactly one intent label.

Use the full query as evidence for your classification -- not
surface wording alone.
Classify the query by the user's primary completion condition:
what they mainly need in order to consider the query resolved.
If multiple subparts are present, choose the single label that
best captures the user's primary need.

## Intent Definitions

QI1 Factoid.  The user wants a specific, verifiable fact
such as a name, number, date, definition, rule, status, or
context-independent yes/no answer.  A short factual answer
fully resolves the query.

QI2 Explanation.  The user wants understanding of a cause,
mechanism, principle, reason, relationship, or underlying
process.  The query is resolved when the user understands why
or how something works.

QI3 Instruction.  The user wants steps, procedures, methods,
or diagnostic actions to perform a task, fix a problem, or
close a gap between expected and actual state.  The query is
resolved when the user can act on the answer.  The subject of
the task must be a concrete, executable procedure -- not an
open-ended personal or interpersonal situation.

QI4 Comparison.  The user wants alternatives evaluated
in order to choose, decide, compare, rank, or receive a
recommendation.  The query is resolved by a grounded judgment
using identifiable, sharable criteria.  This includes queries
asking whether something is good, appropriate, sufficient, or
worth doing -- where the answer requires evaluation against
criteria rather than factual lookup or causal explanation.

QI5 Opinion.  The user wants subjective perspective-sharing,
lived experience, ethical reflection, interpersonal
interpretation, or open-ended discussion.  The query is
resolved through viewpoint exchange rather than factual
lookup, explanation, procedural guidance, or criteria-based
comparison.

## Output (strict JSON)

{ "intent_reasoning":  "step-by-step reasoning",
  "intent":  "QI1" | "QI2" | "QI3" | "QI4" | "QI5" }
```

User prompt

```
<query>
{query}
</query>
```

Figure 7: Full classification prompt for Query Intent (QI1–QI5). Issued via `gpt-4o-mini-2024-07-18` with temperature 0 and structured-output JSON schema in strict mode. The system prompt provides the full taxonomy codebook; the user prompt contains the Stack Exchange post title.

System prompt

You are a web source purpose classifier. Given crawled web page content and its source URL, you assign exactly one purpose label.

Use both the source content and the full URL as evidence for your classification -- not isolated signals alone.
Reason step-by-step based on all available information and choose the label that best captures its primary reader-facing purpose.

Purpose Definitions

SP1 To Promote. The page primarily exists to market, sell, position, or represent a company, product, service, or brand. Its main function is commercial promotion, business representation, or conversion, even if it also contains factual or explanatory material.

SP2 To Inform. The page primarily exists to present factual, descriptive, explanatory, or reference-style information about a topic, concept, entity, or subject. Its main function is to help the reader understand or look up something. The reader's role is to comprehend, not to execute.

SP3 To Instruct. The page primarily exists to help the reader do something by providing steps, procedures, methods, walkthroughs, or other executable guidance. Its main function is to support task completion or skill execution. The reader's role is to follow along and act, not merely to understand.

SP4 To Report. The page primarily exists to report events, developments, announcements, or other time-linked occurrences. Its main function is to tell the reader what happened, what changed, or what was announced. Includes news articles with a byline and publication date, press releases, government announcements, and research press coverage.

SP5 To Discuss. The page primarily exists as a space for exchange among multiple contributors, such as questions, answers, replies, comments, or community problem-solving. Its main function depends on multi-party participation rather than a single authored voice.

SP6 To Opine. The page primarily exists to express a viewpoint, judgment, interpretation, review, editorial stance, or advocacy position. Its main function is subjective evaluation or perspective-sharing rather than neutral information, procedural guidance, event reporting, or multi-party discussion.

Output (strict JSON)

{ "purpose_reasoning": "*step-by-step reasoning*", "purpose": "SP1".."SP6" }

User prompt

<source_url>
{source_url}
</source_url>

<source_content>
{source_content}
</source_content>

Figure 8: Full classification prompt for Source Purpose (SP1–SP6).

System prompt

```
You are a web source domain classifier. Given crawled web
page content and its source URL, you assign exactly one
subject area label.
Use both the crawled content and the full URL structure
(domain, hostname, path) as evidence for your classification
-- not only the example signals listed below.
Reason step-by-step based on all available information.
## Domain Definitions
SD1 Medical/Health. Diseases, treatments, medications, mental
health, nutrition.
SD2 Legal. Laws, regulations, court decisions, legal rights,
compliance.
SD3 Finance. Personal finance, investing, economics,
taxation, banking, insurance, real estate.
SD4 Education. Education, curriculum, student, degree,
scholarship, exam, learning, tuition, admission.
SD5 Science. Natural sciences, mathematics, physics,
chemistry, biology, astronomy.
SD6 Code/Data. Programming, software, data analysis, machine
learning, AI.
SD7 Technical. IT systems, infrastructure, cloud services,
mechanics, electronics.
SD8 Social/Professional. Society, relationships, workplace,
career, parenting, job search.
SD9 Shopping/Travel. Shopping, product reviews, travel,
accommodation.
SD10 Everyday. Daily life, culture, DIY, hobbies, home,
lifestyle, sports, entertainment, pets, food, cooking.
## Output (strict JSON)
{ "domain_reasoning": "step-by-step reasoning", "domain":
"SD1".."SD10" }
```

User prompt

```
<source_url>
{source_url}
</source_url>
<source_content>
{source_content}
</source_content>
```

Figure 9: Full classification prompt for Source Domain (SD1–SD10). YMYL labels (SD1–SD3) follow Google's Search Quality Rater Guidelines (Google, 2025).

System prompt

You are a web source type classifier. Given crawled web page content and its source URL, you assign exactly one structural type label.

Use both the crawled content and the source URL as evidence for your classification -- not only the example signals listed below.
Evaluate ALL type definitions before making a final decision.
Reason step-by-step based on all available information.

Type Definitions

ST1 Official Institution. Content formally issued under the name of: government bodies, legislative institutions, public regulatory agencies; intergovernmental organizations (e.g., UN, WHO, EU, IMF); accredited nonprofits, professional associations, or academic institutions. URL signal examples: .gov, .go.**, .int, .ac.**, .edu.

ST2 Paper/Research. Peer-reviewed academic paper published in a journal or conference proceedings. Must have author name, affiliation, abstract, and references. Excludes theses, preprints, and working papers.

ST3 News/Magazine. News or magazine article published by a media outlet. Must have BOTH a named INDIVIDUAL author (byline) identifiable in the content AND a publication date.

ST4 Wiki/Forum. Content collectively created and maintained by a community. Includes wikis, Q&A platforms, forums, and discussion boards. URL signal examples: wikipedia.org, reddit.com, stackoverflow.com, quora.com.

ST5 Blog/Social. Content created and published by an INDIVIDUAL. Includes blogs, social media posts, and personal channel pages. URL signal examples: twitter.com, x.com, youtube.com, instagram.com, tiktok.com, facebook.com, medium.com, substack.com.

ST6 Private Company. Content published by a private company or non-accredited organization as the publisher. Includes product pages, documentation, and corporate blog posts.

Output (strict JSON)

{ "type_reasoning": "*step-by-step reasoning*", "source_type": "ST1".."ST6" }

User prompt

<source_url>
{source_url}
</source_url>

<source_content>
{source_content}
</source_content>

Figure 10: Full classification prompt for Source Type (ST1–ST6).

System prompt

You are an answer-source fidelity evaluator. Given a cited sentence from an LLM response and the full text of the cited source, you assign exactly one fidelity label.

Use the cited_sentence as the primary evaluation target. Read the ENTIRE source_content before making a judgment.
Identify the cited_sentence's main claim -- the central assertion the sentence is built around -- and use that as the basis for your judgment. Treat incidental details (extra adjectives, illustrative terminology, side phrases) as secondary; they do not by themselves disqualify a label.
Reason step-by-step based on all available information.

Decision Procedure

Decide the final label in two steps.

Step 1 -- Macro verdict. Choose one of:

SUPPORTED. The cited_sentence's main claim is present in and consistent with the source content. Citing one valid aspect of a multi-faceted source is acceptable. Natural summarization that omits details, or restates the source in different words while preserving its meaning, scope, and certainty, is SUPPORTED. Minor extra wording (e.g., an illustrative term, an everyday example, a side phrase) that does not change the main claim's truth value does NOT by itself disqualify SUPPORTED.

DISTORTED. The source discusses the SAME topic as the cited_sentence AND contains a specific passage that the cited_sentence is based on, BUT the cited_sentence's version materially differs from the source in a way that changes the claim's meaning, scope, certainty, or attribution. Use DISTORTED when you can point to the specific source passage being altered. If the alteration is only cosmetic and the main claim still holds, prefer SUPPORTED.

FABRICATED. The cited_sentence's main claim cannot be located in the source at all. This applies when the source covers a different topic entirely, or discusses the same broad topic but never makes -- or even gestures at -- the specific claim asserted. Do NOT use FABRICATED merely because the cited_sentence adds incidental details on top of a claim that IS supported; that case is SUPPORTED (if the extras are minor) or DISTORTED (if the extras materially change the claim).

Step 2 -- Final label.

If SUPPORTED → output **ASF5**. If FABRICATED → output **ASF1**.
If DISTORTED → choose exactly one distortion mechanism below and output the corresponding ASF label.

Distortion Mechanisms (used only when Step 1 = DISTORTED)

ASF4 Amplified. The main claim exists in the source but the cited_sentence presents it with materially greater certainty, scope, or generality. The source uses explicit qualifiers (e.g., "may," "suggests," "in some cases," "preliminary") that the cited_sentence strips away, OR the cited_sentence extends a narrowly-scoped finding into a broader claim, OR the cited_sentence adds a substantive procedural step / specific numerical figure / branded slogan that materially changes what the source actually says.

ASF3 Contradicted. The source concludes or argues the opposite of what the cited_sentence presents.

ASF2 Misattributed. The cited_sentence's main claim is plausible AND the source contains recognizably related content, BUT the claim comes from a different part of the source's discussion or is attributed to a context the source does not actually address.

Tie-breakers

- If the main claim is supported but extras are present: prefer SUPPORTED for minor incidental extras; prefer ASF4 (Amplified) only when the extras materially change the claim.
- If you are torn between SUPPORTED and FABRICATED, ask whether the source contains the main claim at all. If yes → SUPPORTED. If no → consider DISTORTED first, then FABRICATED.

Output (strict JSON)

{ "asf_reasoning": *"step-by-step reasoning"*, "verdict": "ASF1".."ASF5" }

User prompt

<cited_sentence>
{cited_sentence}
</cited_sentence>

<source_content>
{source_content}
</source_content>

Figure 11: Full adjudication prompt for Answer–Source Fidelity (ASF1–ASF5).

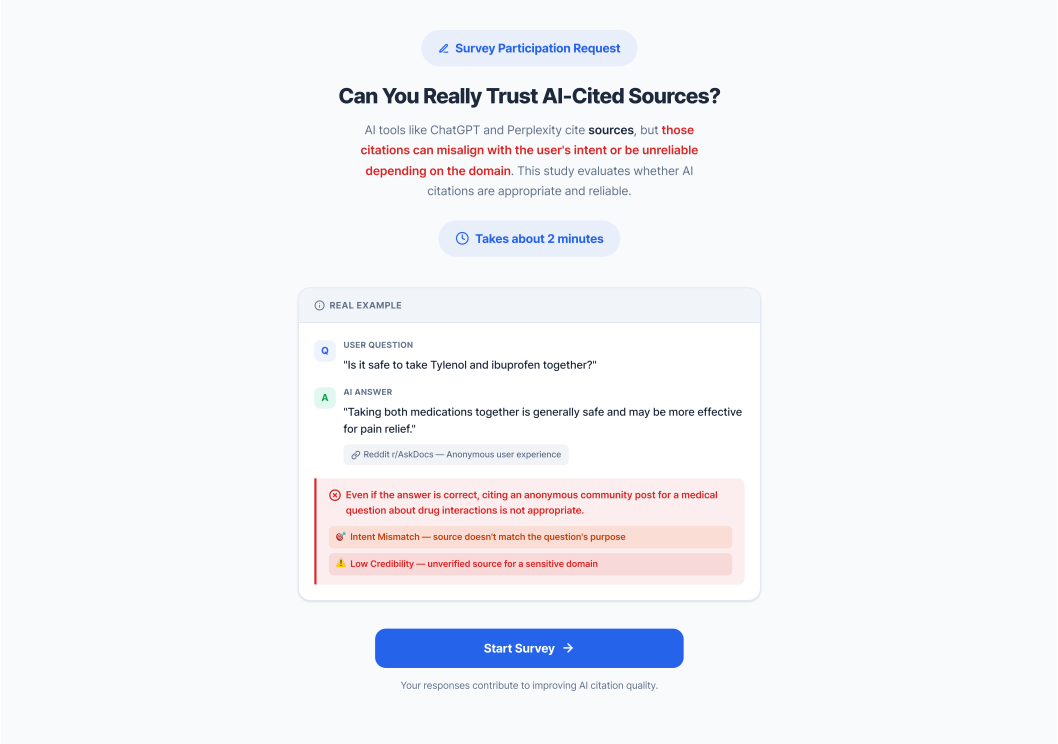

Figure 12: Screenshot of the expert-panel survey landing page used for IPA Matrix validation (§C.1) and SS Matrix validation (§C.2). The page introduces the rating task and presents a real-world example before participants begin.

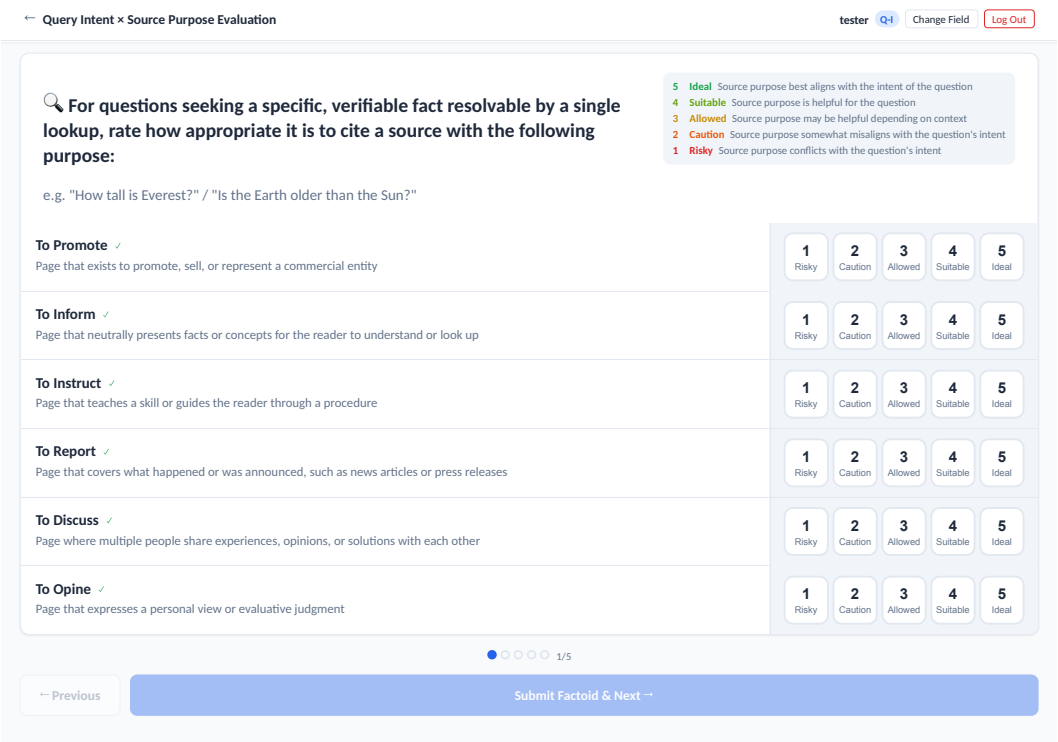

Figure 13: Screenshot of the IPA matrix expert-validation interface (§C.1). On a single page, validators rate the appropriateness of each of the six source purposes (SP1–SP6) for one query intent (here QI1 Factoid) on the 1–5 scale. Validators complete five such pages, one per intent.

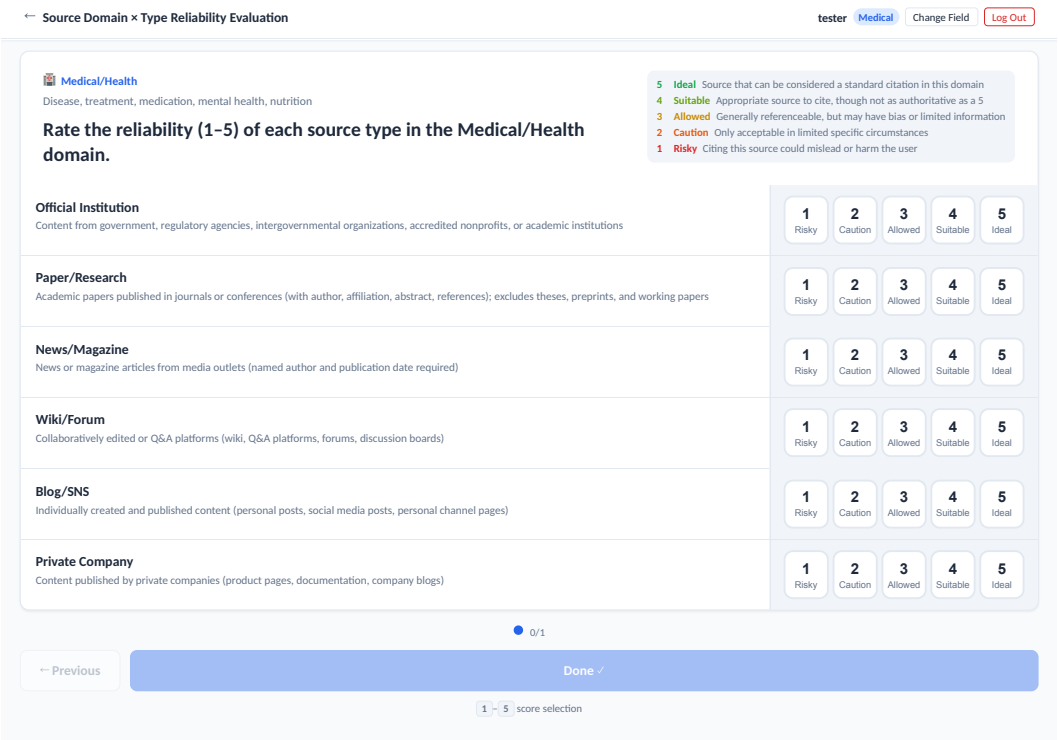

Figure 14: Screenshot of the SSM expert-validation interface (§C.2). On a single page, validators rate the reliability of each of the six source types (ST1–ST6) for one assigned source domain (here SD1 Medical/Health). Each of the ten domains is rated independently by $n = 10$ domain experts.

D THE VM EFFECT: DETAILED EMPIRICAL RESULTS

This appendix expands the empirical results summarized in Section 4. Eight subsections organize the material from the broadest aggregate view to fine-grained robustness checks: structural composition of the source pool (§D.1), aggregate failure rates across the three evaluation dimensions (§D.2), the fidelity–suitability trade-off across models (§D.3), how provider, scale, and reasoning shape citation quality (§D.4), response-level failure exposure (§D.5), robustness and lower bounds (§D.6), external validation on Search Arena (§D.7), and qualitative failure analysis (§D.8). All metrics are computed on the final evaluable pool of 761,495 citation triples (Appendix B.3, Table 11).

D.1 A STRUCTURALLY BIASED SOURCE POOL

Source pool composition. The pool of sources from which models cite is structurally skewed before any dimension-level failure occurs (Table 29). The skew concentrates on what kind of source the models choose (Source Type) and what those sources are written for (Source Purpose); the domains those sources cover are, in contrast, distributed nearly uniformly. The two largest Source Types are Blog/Social (29.2%) and Private Company (21.1%): community-edited and corporate sources together account for about half of all citations (50.3%), nearly twice the 26.3% from Official Institution (17.5%) and Paper/Research (8.8%) combined. Source Purpose is even more concentrated: To Inform (73.1%) and To Instruct (17.1%) together account for 90.2% of all citations, while the four remaining purposes (To Discuss 4.9%, To Promote 2.5%, To Report 1.6%, and To Opine 0.8%) collectively make up only 9.8%. Source Domain, by contrast, is approximately uniform across the seven leading domains, ranging from Technical (14.1%) down to Legal (10.5%) within a four-point spread; the high-stakes YMYL domains, comprising Medical/Health, Legal, and Finance, together accounts for 34.9% of all citations.

D.2 AGGREGATE FAILURE RATES ACROSS THREE DIMENSIONS

The 761,495 evaluable citations admit three distinct failure modes, one per dimension: Intent–Purpose Alignment, Source Suitability, and Answer–Source Fidelity. The aggregate Fidelity Failure Rate (FFR) of 30.6% on Answer–Source Fidelity is the largest of the three; the Suitability Failure Rate (SFR) of 27.1% on Source Suitability is comparable; the Alignment Failure Rate (AFR) of 5.1% on Intent–Purpose Alignment is approximately one-sixth of the FFR.

Answer–Source Fidelity. The fidelity distribution is sharply bipolar: 61.2% Supported, 24.5% Fabricated, 14.3% intermediate. The aggregate FFR is 30.6%, with mean ASF = 3.74 (SD = 1.73). When fidelity fails, it tends to fail outright rather than partially.

Source Suitability. Scores concentrate around Borderline (score 3, 33.3%) but unlike fidelity show no bipolar gap. The failure region (scores 1–2) accounts for 27.1%: Weak Mismatch (21.5%) and Severe Mismatch (5.6%). Most suitability failures are marginal rather than catastrophic—a source one step removed from the ideal type rather than wholly inappropriate. The aggregate SFR (SS $\leq$ 2) is 27.1%, with mean SS = 3.32 (SD = 1.22).

Intent–Purpose Alignment. The alignment distribution is concentrated at score 5 (63.2%), with score 3 as a secondary mode (20.5%) and only 5.1% in the failure region (scores 1–2). Mean IPA = 4.31 (SD = 1.01). The low failure rate reflects the makeup of the source pool: Inform-purpose sources account for 73.1% of all citations and score 4 or 5 against the two most common query intents (Explanation 44.8%, Factoid 18.4%).

Three-dimension summary. The three dimensions differ not only in failure rates but in what drives them (Table 31). Answer–Source Fidelity has the highest failure rate (FFR 30.6%) and the largest model effect ($\eta^2_H = 0.087$, medium); Source Suitability has a slightly lower aggregate rate (SFR 27.1%) but is the most domain-sensitive dimension, with YMYL SFR nearly doubling the non-YMYL rate (Table 32); Intent–Purpose Alignment has the lowest rate (AFR 5.1%) and is the least consequential for aggregate citation quality. The three dimensions capture distinct phenomena rather than a single quality factor, as illustrated by the cross-dimension analysis in §D.6.

Table 29: Source pool composition across the 761,495 evaluable citations, broken down by Source Type, Source Purpose, and Source Domain. Inline bars (■ Sage Green; 1% = 1.2 pt) visualize each share. Highlighted cells mark the dominant labels in Source Type (Blog and Company, which together account for 50.3% of citations) and in Source Purpose (Inform and Instruct, which together account for 90.2%). Source Domain remains otherwise unhighlighted because its distribution is approximately uniform across the leading domains. The **YMYL** column marks the three high-stakes Source Domains (Medical, Legal, and Finance) with ✓, and their rows are highlighted accordingly.

Dimension	Label	YMYL	Count	%	Share
Source Type	**Blog/Social**		**222,151**	**29.2**	
	Company		**160,584**	**21.1**	
	Official		133,495	17.5	
	Wiki/Forum		103,095	13.5	
	News		75,264	9.9	
	Research		66,906	8.8	
Source Purpose	**To Inform**		**556,903**	**73.1**	
	To Instruct		**130,329**	**17.1**	
	To Discuss		36,980	4.9	
	To Promote		18,953	2.5	
	To Report		12,176	1.6	
	To Opine		6,154	0.8	
Source Domain	Technical		107,179	14.1	
	Code/Data		106,155	13.9	
	Medical	✓	102,686	13.5	
	Science		94,770	12.4	
	Everyday		88,100	11.6	
	Finance	✓	82,881	10.9	
	Legal	✓	79,881	10.5	
	Social		43,473	5.7	
	Shopping		28,369	3.7	
	Education		28,001	3.7	

Table 30: Score distribution across the three citation-quality dimensions (761,495 evaluable citations). Answer–Source Fidelity is sharply bipolar (Supported 61.2%, Fabricated 24.5%, middle nearly empty). Source Suitability concentrates on Borderline (33.3%). Intent–Purpose Alignment is dominated by score 5 (63.2%). Inline bars (■ Amber for ASF, ■ Sage for SS, ■ Steel for IPA; 1% = 0.6 pt) visualize each share.

Dimension	Label	Count	%	Share
ASF (Fidelity)	Supported (5)	465,844	61.2	
	Amplified (4)	54,396	7.1	
	Contradicted (3)	8,269	1.1	
	Misattributed (2)	46,119	6.1	
	Fabricated (1)	186,867	24.5	
SS (Suitability)	Suitable (5)	188,217	24.7	
	Adequate (4)	112,862	14.8	
	Borderline (3)	253,898	33.3	
	Inadequate (2)	163,815	21.5	
	Unsuitable (1)	42,703	5.6	
IPA (Alignment)	Structural Match (5)	481,054	63.2	
	Functional Support (4)	85,230	11.2	
	Partial Relevance (3)	156,419	20.5	
	Weak Fit (2)	27,355	3.6	
	Structural Conflict (1)	11,437	1.5	

Table 31: Three-dimension summary of aggregate means and effect sizes. The contrast between Answer–Source Fidelity (model-driven, $\eta^2 = 0.087$) and Intent–Purpose Alignment (category-driven, $\eta^2_{\text{cat}} = 0.041$) supports treating the three dimensions as distinct failure modes rather than a single composite quality score.

Dimension	Mean (1–5)	Failure rate	Model η^2	Category η^2
Answer–Source Fidelity	**3.74** ($\pm$1.73)	FFR 30.6%	**0.087** (medium)	0.002 (negligible)
Source Suitability	**3.32** ($\pm$1.22)	SFR 27.1%	0.041 (small)	0.026 (small)
Intent–Purpose Alignment	**4.31** ($\pm$1.01)	AFR 5.1%	0.003 (negligible)	**0.041** (small)

Table 32: YMYL versus non-YMYL Suitability Failure Rate. YMYL domains (Medical, Legal, and Finance) have 1.81$\times$ the SFR of non-YMYL domains (odds ratio = 2.317).

Group	n	R-SS	SFR %
YMYL (Medical + Legal + Finance)	265,448	3.226	**38.3**
non-YMYL	496,047	3.363	21.1

D.3 The Fidelity–Suitability Trade-off Across Models

Answer–Source Fidelity is model-driven. Disaggregating to the model level reveals a 3.6-fold FFR spread: claude-haiku at 12.3% versus gpt-5-mini at 44.8% (Figure 15). A Kruskal–Wallis decomposition assigns $\eta^2_H = 0.087$ (medium effect; $H = 66{,}265$, $p < 0.001$) to model identity but only $\eta^2_H = 0.002$ (negligible) to query category, a 46-fold ratio (Table 31). Fidelity failure is driven almost entirely by model identity rather than by query category.

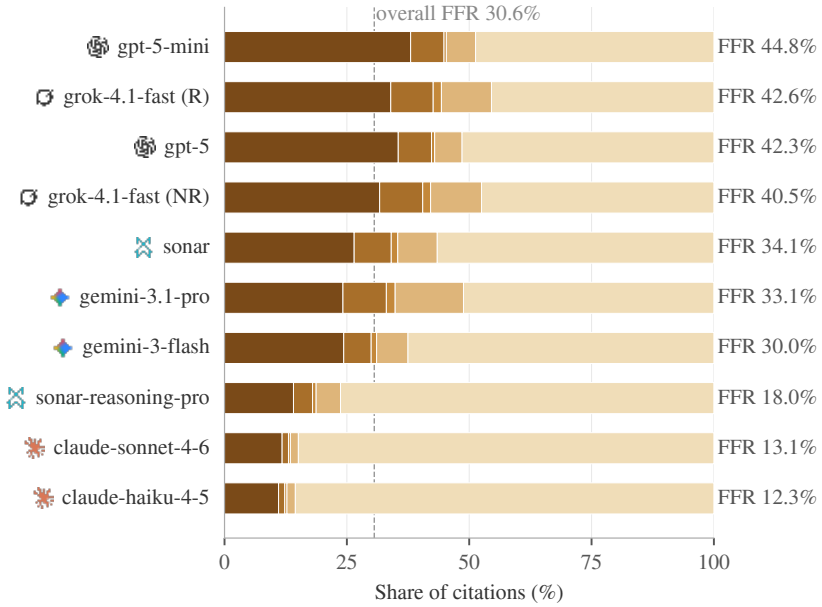

Figure 15: Per-model Answer–Source Fidelity score distribution across the 761,495 evaluable citations, sorted by Fidelity Failure Rate (FFR, scores ≤ 2). Each horizontal bar decomposes into five rubric levels in a Warm Amber gradient: Fabricated (■), Misattributed (■), Contradicted (■), Amplified (■), and Supported (■). The dashed vertical line marks the aggregate FFR (30.6%); right-margin labels report per-model FFR. The 3.6-fold spread—from claude-haiku (12.3%) to gpt-5-mini (44.8%)—is the largest cross-model gap among the three evaluation dimensions.

Source Suitability partly inverts the fidelity ranking. SFR ranges from 8.0% (gpt-5) to 33.5% (gemini-3-flash), a 4.2-fold spread (Table 33). The Kruskal–Wallis model effect ($\eta^2_H = 0.041$, small) is weaker than for fidelity but still exceeds the category effect ($\eta^2_H = 0.026$): source suitability, too, is model-determined. Yet the ranking partly inverts the fidelity ranking: the Anthropic models, top-ranked on FFR (ranks 1–2), drop to SFR ranks 7 and 9, while gpt-5 holds SFR rank 1 despite FFR rank 8. The rank shift Δ ranges from $+8$ (gpt-5-mini) to -7 (claude-sonnet), with every model above the inversion boundary exhibiting a corresponding SFR decline. This inversion sharpens further in YMYL domains (§D.2, Table 32).

Good citers pick bad sources. The fidelity–suitability inversion traces to source selection: models with low FFR systematically draw from less authoritative source types (Table 34). gpt-5 concentrates 41.8% of its citations on Official sources and achieves the lowest SFR (8.0%); the Anthropic models route 32–35% into Blog and only 12% into Official, producing SFR above 30%. Company share remains near 21% across all models, so the cross-model SFR gap is driven almost entirely by the Official–Blog axis. Figure 16 plots each model's FFR against its SFR: no model approaches the ideal corner, and models with similar source-type profiles occupy similar positions along the iso-failure contours, suggesting that the trade-off arises from the interplay between retrieval infrastructure and model-level source selection rather than from any single factor. The gap widens in YMYL domains (§D.2, Table 32).

Table 33: Fidelity–suitability rank inversion across ten models. Models are sorted by FFR (ascending, best first); SFR rank is computed independently. The Δ column reports the rank shift (FFR rank − SFR rank): a ▼ indicates the model's suitability rank is worse than its fidelity rank; ▲ indicates the reverse. Inline bars (■ Amber FFR, ■ Sage SFR; 1 % = 0.5 pt) visualize each rate.

Model	R_F	FFR %	R_S	SFR %	Δ
claude-haiku-4-5	**1**	12.31	7	30.07	▼ 6
claude-sonnet-4-6	**2**	13.07	9	31.65	▼ 7
sonar-reasoning-pro	**3**	18.00	8	31.48	▼ 5
gemini-3-flash	4	29.97	10	33.46	▼ 6
gemini-3.1-pro	5	33.07	6	30.07	▼ 1
sonar	6	34.07	5	29.99	▲ 1
grok-4.1-fast (NR)	7	40.50	4	24.79	▲ 3
gpt-5	8	42.29	**1**	7.98	▲ 7
grok-4.1-fast (R)	9	42.64	3	24.56	▲ 6
gpt-5-mini	10	44.85	**2**	14.25	▲ 8

Table 34: Per-model Source Type distribution (% of each model's citations), sorted by SFR ascending (best first). Each row sums to 100%. Inline bars highlight the key contrast: ■ Sage for Official, ■ Amber for Blog. Models with higher Official shares exhibit lower SFR; models routing more citations into Blog exhibit higher SFR. The Spearman correlation between Official share and SFR is $\rho = -0.84$.

Model	R_S	Offi.	Blog	Res.	Comp.	Wiki	News
gpt-5	1	**41.82**	5.34	17.56	21.23	7.29	6.77
gpt-5-mini	2	31.19	13.01	17.66	22.87	8.75	6.52
grok-4.1-fast (R)	3	19.27	26.01	6.51	19.02	17.38	11.81
grok-4.1-fast (NR)	4	19.81	26.68	7.08	19.99	15.21	11.22
sonar	5	15.75	33.67	6.87	21.78	12.67	9.26
gemini-3.1-pro	6	12.29	32.86	7.64	20.52	16.42	10.26
claude-haiku-4-5	7	12.76	31.82	9.51	21.92	13.31	10.68
sonar-reasoning-pro	8	13.32	**36.24**	6.84	21.26	12.19	10.13
claude-sonnet-4-6	9	11.78	34.50	10.06	20.65	12.79	10.22
gemini-3-flash	10	**13.28**	33.31	4.83	24.44	15.10	9.04
Total		17.53	29.17	8.79	21.09	13.54	9.88

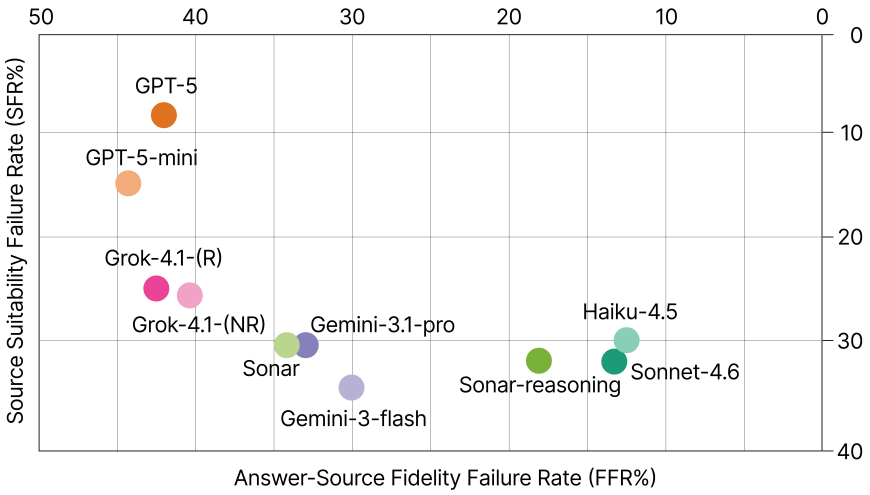

Figure 16: Fidelity–suitability trade-off across ten models. Each bubble plots a model's Fidelity Failure Rate (FFR; x-axis, reversed) against its Suitability Failure Rate (SFR; y-axis, reversed); the ideal corner (0 %, 0 %) lies at the upper right, but no model approaches it. Color encodes provider: models sharing a search backend cluster together, consistent with the 88–96 % between-provider variance share reported in §D.4.

D.4 How Provider, Scale, and Reasoning Shape Citation Quality

Providers exhibit structurally distinct profiles. The five providers do not align on a single "citation quality" dimension but trade off across the three dimensions in structurally different ways (Table 35). Anthropic exhibits FFR = 12.9% and SFR = 31.3% (best fidelity, worst suitability); OpenAI exhibits FFR = 43.8% and SFR = 11.7% (worst fidelity, best suitability). Google and Perplexity cluster near the pool-wide mean on both dimensions, while xAI pairs high FFR (41.6%) with moderate SFR (24.7%). The Anthropic–OpenAI inversion documented at the model level in §D.3 thus extends to the provider level, consistent with each provider operating a distinct search infrastructure. The underlying score distributions (Figure 17) confirm that these rate differences reflect distinct distributional shapes rather than uniform shifts: Anthropic's fidelity scores concentrate at 5 while OpenAI's are bimodal, and the pattern reverses on suitability.

Table 35: Provider-level three-dimension failure profile. Anthropic and OpenAI exhibit nearly inverted FFR/SFR rankings; no provider excels on all three dimensions. Inline bars (■ Amber FFR, ■ Sage SFR, ■ Steel AFR; 1 % = 0.5/0.5/4 pt respectively).

Provider	n	Share	FFR	SFR	AFR
Anthropic	167,078	21.9	**12.9**	31.3	4.3
OpenAI	93,250	12.2	43.8	**11.7**	3.8
Perplexity	191,568	25.2	28.2	30.5	5.8
Google	126,328	16.6	31.9	31.4	4.9
xAI	183,271	24.1	41.6	24.7	5.9

Provider identity dominates quality variance. A two-way ANOVA decomposing per-dimension means by provider (between) and within-provider model confirms this pattern quantitatively (Table 36): provider effects account for 96.5% of IPA variance and 96.1% of SS variance, leaving under 4% for within-provider model differences. ASF variance is also provider-dominated at 88.3%, but the remaining 11.7% is attributable to within-provider model differences—roughly three times the residual of the other two dimensions. The asymmetry has a structural explanation: IPA and SS are determined by the retrieval backend alone, which fixes the source's purpose and type before the generator sees them, so models sharing a backend score nearly identically. ASF depends also on how

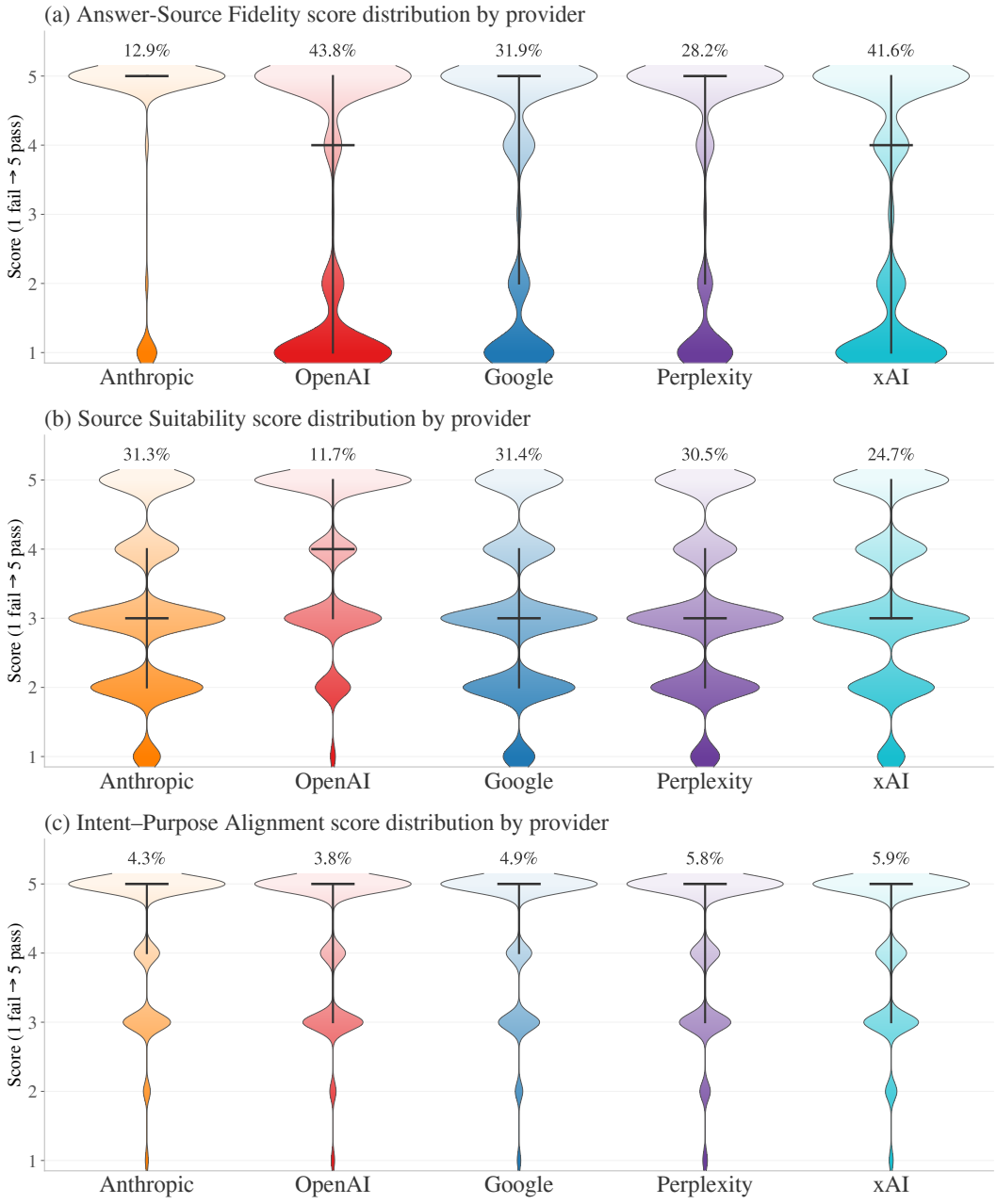

Figure 17: Score distributions by provider across three dimensions. (a) Answer–Source Fidelity: Anthropic concentrates at score 5 while OpenAI and xAI show bimodal distributions with substantial mass at score 1. (b) Source Suitability: the pattern partly inverts, with OpenAI concentrates at higher scores while Anthropic and Google spread toward score 3. (c) Intent–Purpose Alignment: all five providers show near-identical distributions concentrated at score 5, confirming that alignment is not provider-driven.

well the generator utilizes retrieved content, giving model-level language ability room to affect the outcome that the other two dimensions do not permit.

Table 36: Two-way ANOVA decomposition of per-dimension variance into provider effects (between providers) and model effects (within provider). Provider effects dominate all three dimensions, accounting for 88–96% of total variance.

Metric	$\text{SumSq}_{\text{btw}}$	%	$\text{SumSq}_{\text{within}}$	%
IPA score	1,993.6	**96.5**	72.3	3.5
AS score	152,431.1	**88.3**	20,222.1	11.7
SS score	44,370.0	**96.1**	1,823.2	3.9

Larger models do not improve citation quality. Within-provider model pairs that differ in scale but share a retrieval backend show marginal quality differences (Table 37). The Anthropic pair (sonnet versus haiku) differs by less than 0.8 pp on FFR and 1.6 pp on SFR despite sonnet issuing roughly twice as many citations per response. The OpenAI pair (gpt-5 versus gpt-5-mini) shows the largest within-provider gap: gpt-5 cites less and achieves both lower FFR (-2.6 pp) and lower SFR (-6.3 pp). The Google pair (pro versus flash) reproduces the D.3 trade-off in miniature: pro improves SFR by 3.4 pp but worsens FFR by 3.1 pp. In all three cases the within-provider Δ remains far smaller than the between-provider gaps of Table 35, confirming the 4–12% within-provider share reported above.

Table 37: Within-provider paired comparison of larger versus smaller models sharing the same retrieval backend. All three pairs exhibit $|\Delta\text{FFR}| \leq 3.1$ pp and $|\Delta\text{SFR}| \leq 6.3$ pp, consistent with the 4–12 % within-provider variance share in Table 36. Inline bars (■ Amber FFR, ■ Sage SFR; 1 % = 0.5 pt). Bold marks improved metric within each pair.

Provider	Model	n	Cit/Q	FFR %	SFR %
Anthropic	claude-sonnet-4-6	129,936	12.7	13.07	31.65
	claude-haiku-4-5	37,142	5.3	**12.31**	**30.07**
	Δ (sonnet − haiku)			+0.76	+1.58
OpenAI	gpt-5	38,322	5.8	42.29	**7.98**
	gpt-5-mini	54,928	6.4	44.85	14.25
	Δ (5 − mini)			**−2.55**	**−6.27**
Google	gemini-3.1-pro	77,999	12.7	33.07	**30.07**
	gemini-3-flash	48,329	7.2	**29.97**	33.46
	Δ (pro − flash)			+3.10	**−3.39**

Reasoning improves only Answer–Source Fidelity. The search-versus-generation decomposition can be tested directly using paired reasoning and non-reasoning models that share a retrieval backend (Table 38). The xAI pair (grok-4.1-fast (R) versus grok-4.1-fast (NR)) shows differences of at most 2.2 pp on all three dimensions—no measurable reasoning effect. The Perplexity pair (sonar-reasoning-pro versus sonar) shows a sharper pattern: the reasoning model issues half as many citations (6.8 versus 11.4) while reducing FFR by 16.1 pp (18.0% versus 34.1%) and raising mean ASF by 0.65. But SFR and AFR differ negligibly (Δ SFR = +1.5 pp, Δ AFR = -0.3 pp). When reasoning helps, it helps Answer–Source Fidelity only; source selection is governed by the search backend and is beyond the reach of generation-side improvements.

D.5 Response-Level Failure Exposure

Response-level failure exposure. The metrics reported above are citation-level rates; the user-facing impact is at the response level, where any single failed citation constitutes user exposure (Table 39). For a model averaging $\bar{n}$ citations per response, even a modest citation-level rate

Table 38: Reasoning versus non-reasoning paired comparison within xAI and Perplexity. The Perplexity pair shows substantial improvement on Answer–Source Fidelity (Δ FFR = −16.1 pp) but negligible change on Source Suitability (Δ SFR = +1.5 pp). The xAI pair shows no effect on either dimension. Inline bars (■ Amber FFR, ■ Sage SFR; 1 % = 0.5 pt). Bold marks improved metric within each pair.

Provider	Model	n	Cit/Q	FFR %	SFR %
xAI	grok-4.1-fast (R)	92,355	9.0	42.64	**24.56**
	grok-4.1-fast (NR)	90,916	8.8	**40.50**	24.79
	Δ (R − NR)			+2.13	−0.22
Perplexity	sonar-reasoning-pro	69,411	6.8	**18.00**	31.48
	sonar	122,157	11.4	34.07	**29.99**
	Δ (reas. − sonar)			**−16.07**	+1.49

compounds rapidly. Claude-sonnet's FFR is only 13.1%, but with $\bar{n} = 12.7$ its response-level exposure (R-FFR) reaches 62.9%; its SFR of 31.7% produces R-SFR of 70.6%. The amplification is universal: the best-performing model (claude-haiku, $\bar{n} = 5.3$) still exposes 71.3% of its responses to at least one failure on any axis (Any), and models averaging $\bar{n} \geq 9$ have Any rates of at least 89.5%. The "good citers, bad sources" trade-off of §D.3 is not diluted but amplified at the response level: claude-sonnet achieves the third-lowest R-FFR (62.9%) yet the highest R-SFR (70.6%).

Table 39: Citation-level failure rates versus empirical response-level exposure across ten models, sorted by Any descending. $\bar{n}$ is the mean number of evaluable citations per response. R-FFR, R-SFR, and R-AFR report the share of responses containing at least one citation that fails on the corresponding dimension; Any reports responses with at least one failure on any dimension. Even the best-performing model (claude-haiku, $\bar{n} = 5.3$) exposes 71.3% of its responses to at least one failure; models averaging $\bar{n} \geq 9$ reach 89–96%. Inline bars (■ Amber R-FFR, ■ Sage R-SFR, ■ Steel R-AFR, ■ Black Any; 1 % = 0.22 pt).

Model	$\bar{n}$	FFR	R-FFR	SFR	R-SFR	AFR	R-AFR	Any
grok-4.1 (R)	9.0	42.6	91.8	24.6	54.2	6.0	19.2	**96.1**
grok-4.1 (NR)	8.8	40.5	90.1	24.8	57.3	5.8	19.4	**95.0**
gemini-3.1-pro	12.7	33.1	87.8	30.1	54.4	5.2	15.5	**94.0**
sonar	11.4	34.1	87.1	30.0	55.6	5.9	15.9	**94.0**
gpt-5-mini	6.4	44.8	85.7	14.2	36.7	4.1	12.3	**89.7**
claude-sonnet	12.7	13.1	62.9	31.7	70.6	4.3	17.4	**89.5**
gpt-5	5.8	42.3	80.9	8.0	21.9	3.5	10.2	**84.7**
gemini-3-flash	7.2	30.0	64.8	33.5	56.3	4.3	12.0	**82.7**
sonar-reas.	6.8	18.0	56.4	31.5	49.1	5.5	12.1	**79.2**
claude-haiku	5.3	12.3	35.6	30.1	51.8	4.4	10.1	**71.3**

D.6 Robustness and Lower Bounds

The three dimensions capture distinct failure modes. The observed CritVM rate is 3,174 of 761,495 citations (0.42%), close to the expectation under independence, AFR · FFR · SFR = $0.051 \cdot 0.306 \cdot 0.271 \approx 0.0042$; both rates round to 0.42%. The dimensions also show different patterns of association with model identity and query category. Answer–Source Fidelity is model-driven ($\eta^2_H = 0.087$, medium); Source Suitability is model-driven but amplified by YMYL domains ($\eta^2_H = 0.041$, OR = 2.32; Table 32); Intent–Purpose Alignment is category-driven ($\eta^2_{\text{cat}} = 0.041$), with Science queries achieving R-IPA 4.60–4.70 regardless of model while Professional queries fall to 4.00–4.47. The Venn decomposition illustrates the distinct failure patterns: ASF-only failures account for 22.2% of all citations, SS-only for 19.8%, and IPA-only for 2.1%, while the three-dimension intersection is 0.42%—at least 5 times smaller than even the smallest single-dimension bucket.

Threshold sensitivity. Across the baseline and six ± 1 perturbation variants of the ≤ 2 failure threshold, model-level CritVM rankings preserve Kendall $\tau \geq 0.82$ in five of seven settings (Table 40). The two exceptions are ss_strict (≤ 1, $\tau_{\text{model}} = 0.733$) and ipa_strict ($\leq 1$, $\tau_{\text{model}} = 0.644$), where the binary region is too sparse for stable rankings. Loosening the ASF threshold (asf_loose = ≤ 3) preserves rankings most strongly ($\tau = 0.956$); tightening ASF (asf_strict = ≤ 1) preserves $\tau = 0.911$. The reported threshold is a stable operating point.

Table 40: Threshold sensitivity: the baseline and six ± 1 perturbation variants. Five of seven settings preserve $\tau_{\text{model}} \geq 0.82$. The reported ≤ 2 threshold is a stable operating point.

Variant	IPA$\leq$	ASF$\leq$	SS$\leq$	n CritVM	%	τ_{model}	τ_{cat}
baseline	**2**	**2**	**2**	**3,174**	**0.417**	**1.000**	**1.000**
as_loose	2	3	2	3,278	0.430	0.956	1.000
as_strict	2	1	2	2,593	0.341	0.911	0.867
ipa_loose	3	2	2	14,290	1.877	0.867	0.600
ss_loose	2	2	3	12,407	1.629	0.822	0.200
ss_strict	2	2	1	744	0.098	0.733	0.067
ipa_strict	1	2	2	1,006	0.132	0.644	−0.200

Temporal stability. Splitting the 15-day collection window at the median date (April 3, 2026) yields balanced per-dimension means with ≤ 5 pp shift; no model changes its rank-quartile assignment between halves (Table 41). The reported aggregate metrics are not the artifact of a single news cycle.

Table 41: Temporal stability across the 15-day collection window. Splitting at the median date (April 3, 2026) yields balanced per-dimension means with ≤ 5 pp shift. Inline bars (■ Amber FFR, ■ Sage SFR, ■ Steel AFR; 1% = 0.6/0.6/4 pt respectively).

Half	n	FFR	SFR	AFR
First (3/26–4/02)	106,363	34.1	24.8	4.7
Second (4/03–4/09)	655,132	30.0	27.5	5.2

Table 42: Phantom-citation and PDF-citation rates by model. Phantom citations (URL never reachable) and PDF citations fail at higher rates in the crawl pipeline; both bias reported failure rates downward. Black inline bars (1% = 1.7 pt) visualize phantom rates.

Model	Provider	Total	Phantom	%	PDF	PDF %
gemini-3.1-pro	Google	189,166	29,313	**15.50**	13,895	7.35
gemini-3-flash	Google	117,218	16,923	14.44	13,119	11.19
claude-haiku-4-5	Anthropic	48,249	2,416	5.01	1,616	3.35
grok-4.1-fast (NR)	xAI	184,872	7,674	4.15	15,302	8.28
claude-sonnet-4-6	Anthropic	160,217	5,764	3.60	5,616	3.51
gpt-5-mini	OpenAI	76,713	2,366	3.08	10,178	13.27
gpt-5	OpenAI	55,337	1,491	2.69	9,306	**16.82**
grok-4.1-fast (R)	xAI	173,077	4,088	2.36	11,505	6.65
sonar	Perplexity	177,185	3,994	2.25	27,210	15.36
sonar-reasoning-pro	Perplexity	89,012	1,897	2.13	11,845	13.31

Crawl failures bias toward underestimation. Of the 1,271,046 raw citation pairs, 36.8% failed to crawl (Appendix B.3, Table 8), and these failures are concentrated on Forum/Q&A (67.9% failure rate) and Blog/Social (55.0%) host tiers. Both tiers map onto SS cells with values ≤ 2 in YMYL domains. A second underestimation source is phantom citations: Google's two models exhibit phantom-citation rates of 14.4–15.5% versus 2.1–5.0% for other models (Table 42), where the cited

URL was never reachable; treating these as evaluation-eligible would worsen Google's measured SFR and FFR. A third source is PDF-format citations: gpt-5 (16.8%) and sonar (15.4%) cite PDFs at substantially higher rates, and PDF failures are concentrated on Research and Official sources whose absence inflates measured SFR. The FFR, SFR, and AFR values reported throughout this appendix should be read as conservative lower bounds on the true population values.

D.7 External Validation on Search Arena

Setup. We evaluated the framework on Search Arena queries to examine whether the main findings extend beyond Stack Exchange under different query distributions. From 24,069 conversations in Search Arena (Miroyan et al., 2026), we retained first-turn English queries of 15–150 characters and removed duplicates, leaving 6,258 candidates. We sampled 100 queries from each of Search Arena's nine intent categories to obtain 900 queries. We applied the same prompts, crawling pipeline, judge, and scoring matrices used in the Stack Exchange evaluation. The experiment covered eight models from four providers because the two xAI models were unavailable through the API at collection time.

Results. The fidelity–suitability contrast between Anthropic and OpenAI also appeared on Search Arena. As shown in Table 43, both Anthropic models had lower FFR but higher SFR than both OpenAI models on both datasets. Absolute failure rates varied across datasets, but the Anthropic–OpenAI contrast remained consistent across the two query sources.

Table 43: Citation-level failure rates on Stack Exchange (SE) and Search Arena.

	FFR (%)		SFR (%)	
Model	**SE**	**Arena**	**SE**	**Arena**
claude-sonnet-4-6	13.1	8.5	31.7	29.7
claude-haiku-4-5	12.3	9.2	30.1	27.6
gemini-3-flash	30.0	34.9	33.5	24.7
sonar-reasoning-pro	18.0	39.2	31.5	28.0
sonar	34.1	39.4	30.0	24.7
gpt-5	42.3	41.1	8.0	8.7
gpt-5-mini	44.8	41.4	14.2	13.0
gemini-3.1-pro	33.1	41.4	30.1	26.8

D.8 Qualitative Failure Analysis

Overview. Six cases drawn from the CiteTrace corpus illustrate distinct failure patterns at the citation and query levels. Case 1 illustrates Critical VM (CritVM), with failures on all three dimensions. Case 2 presents a citation that passes all three dimensions. Case 3 shows that high Answer–Source Fidelity can coexist with low Source Suitability. Cases 4 and 5 compare responses to the same query across models and providers, highlighting differences in Answer–Source Fidelity and source selection. Case 6 presents a citation whose source content could not be retrieved, preventing evaluation of the corresponding citation triple across the three dimensions.

Case 1: Commercial probiotic blog cited as medical evidence

Query ID Q09021

Query *Does increasing one's probiotic uptake lower the chance of getting sick after swimming in waters of questionable quality?*

Site Skeptics · Culture & Recreation

Model perplexity_sonar (Perplexity)

QI QI2 Explanation

Citing sentence
(cit 1071750):

> *"Commercial sources claim probiotics help resist harmful germs from accidentally ingested contaminated water (e.g., during boil water notices), potentially reducing diarrhea risk or severity. Similarly, for swimming pools or natural waters, daily spore probiotics and prebiotics are recommended to diversify gut flora and strengthen defenses against invaders like those slipping past chlorine."*

Source `https://www.essentialprobiotics.com/probiotics-pool-party/`

Source type SP1 Promote · SD1 Medical · ST5 Blog/Social · Host: Commercial/Other

IPA	ASF	SS	CritVM
1	1	1	✓

Source excerpt:

> *"These chemicals, unfortunately, are very toxic to a person's health. Although they assist in keeping the bad bacteria out of the pool, they also eliminate the good bacteria… 'toxic' does not always mean detrimental or deadly, sometimes it can mean that it is harmful in large doses, such as with chlorine and bromine."*

Verification:
Keyword search on the source text (2,061 chars) for six core terms—spore probiotic, boil water, harmful germs, diversify gut, slipping past chlorine, diarrhea—returned **0/6 matches**. The source discusses chlorine and bromine removing iodine from the body; the citing sentence's specific claims do not appear in the source.

Case 2: Correct citation—IRS official source

Query ID Q05620

Query *IRS notice of owing money for income tax and interest*

Site Personal Finance & Money · Life & Arts

Model perplexity_sonar (Perplexity)

QI QI1 Factoid

Citing sentence
(cit 666090):

> *"Pay in full or request installment agreement, Offer in Compromise, or temporary delay (call number on notice or 800-829-1040; may need Form 433-F/A/B)."*

Source `https://www.irs.gov/businesses/small-businesses-self-employed/temporarily-delay-the-collection-process`

Source type SP2 Inform · SD2 Legal · ST1 Official · Host: Gov/Org

IPA	ASF	SS	CritVM
5	5	5	—

Source excerpt:

> *"Prior to approving your request to delay collection, we may ask you to complete a Collection Information Statement (Form 433-F, Form 433-A, or Form 433-B) and provide proof of your financial status… Call 800-829-1040. Or call the phone number on your bill or notice."*

Verification:
Keyword search for four terms—Form 433-F, Form 433-A, Form 433-B, 800-829-1040—returned **4/4 matches**. All three dimensions score 5: an ideal citation where the source is authoritative, topically aligned, and faithfully reflected.

Case 3: Accurate citation of an inappropriate source—fitness-app blog as medical evidence

Query ID Q04569

Query *Why I should avoid exercising after blood donation?*

Site Medical Sciences · Science

Model claude-haiku-4-5 (Anthropic)

QI QI2 Explanation

Citing sentence
(cit 531985):

> *"You're more likely to experience dizziness, fatigue or feeling faint (a recipe for injury), as well as increasing the risk of bleeding, bruising or soreness."*

Source `https://sweat.com/blogs/wellbeing/can-you-work-out-after-giving-blood`

Source type SP2 Inform · SD1 Medical · ST5 Blog/Social · Host: Commercial/Other

IPA	ASF	SS	CritVM
5	5	1	—

Source excerpt:

> *"What happens if you exercise after giving blood? Basically, you're more likely to experience dizziness, fatigue or feeling faint (a recipe for injury), as well as increasing the risk of bleeding, bruising or soreness."*

Verification:
Keyword search (6,717 chars) for six terms—dizziness, fatigue, feeling faint, bleeding, bruising, soreness—returned **6/6 matches**. The citing sentence nearly mirrors the source verbatim (ASF = 5), but a fitness-app marketing blog is structurally inappropriate for a Medical Sciences query (SS = 1, Medical × Blog/Social cell). The pattern matches the Anthropic provider profile (FFR 12.9%, SFR 31.3%).

Case 4: Same query, ten models—fish-oil Vitamin A

Query ID Q04788

Query *How much Vitamin A is in 1g of Fish Oil?*

Site Medical Sciences · Science

QI QI1 Factoid

Context YMYL query on fish-oil supplements and hypervitaminosis A risk. 10 models, 48 evaluable citations.

Per-model citation summary(this query only):

Model	Provider	Cit.	FFR %	mean ASF
claude-sonnet-4-6	Anthropic	5	0.0	5.00
claude-haiku-4-5	Anthropic	1	0.0	5.00
gpt-5	OpenAI	3	33.3	3.67
gpt-5-mini	OpenAI	1	100.0	1.00
grok-4.1-fast (NR)	xAI	8	37.5	3.25
grok-4.1-fast (R)	xAI	10	70.0	2.30
gemini-3.1-pro	Google	5	80.0	2.40
gemini-3-flash	Google	2	100.0	1.00
sonar	Perplexity	8	87.5	1.62
sonar-reasoning-pro	Perplexity	5	80.0	1.80

Analysis:

All ten models reach the same conclusion (standard fish-oil supplements contain negligible Vitamin A; cod liver oil should not be conflated), yet the citations diverge sharply. For this query, Anthropic's two models achieve ASF-mean 5.00 (FFR 0%), while Google and Perplexity fall to mean ASF 1.00–2.40 (FFR 80–100%). Here, FFR is the fraction of each model's evaluable citations for this query with ASF ≤ 2. This contrast is consistent with the direction of the aggregate provider-level FFR gap (Anthropic 12.9% versus Perplexity 28.2% across the full dataset).

Case 5: Same YMYL query, provider divergence—VIX futures bounds

Query ID Q08172

Query *VIX future's lower and upper bounds*

Site Quantitative Finance · Business

QI QI1 Factoid

Context Theoretical no-arbitrage bounds on VIX futures. 5 providers, 8 model responses, 25 evaluable citations.

Per-provider citation profile:

Provider	Cit.	Official+Res.	Blog	Company	Mean SS
Anthropic	13	10 (76.9%)	3 (23.1%)	0	4.31
OpenAI	3	2 (66.7%)	0	1 (33.3%)	4.33
Google	1	0	0	1 (100%)	3.00
xAI	5	0	0	5 (100%)	3.00
Perplexity	3	0	3 (100%)	0	2.00

Analysis:

Provider-mean SS differs by approximately a factor of two (OpenAI 4.33 versus Perplexity 2.00). Anthropic and OpenAI cite Official and Research sources for at least two-thirds of their citations, while xAI cites only Company pages and Perplexity cites only blogs. The provider-level pattern identified in the SS variance decomposition of §D.4 (96.1% between providers) is also visible on this single Finance-domain query.

Case 6: Phantom citation—decommissioned government climate resource

Query ID Q03871

Query *What regions of the US will have improved climate given current predictions of climate change?*

Site Earth Science · Science

Model gpt-5 (OpenAI)

QI QI2 Explanation

Citing sentence
(cit 444789):

> *"Northern New England and interior Northeast (VT, NH, ME, Adirondacks/upstate NY)—Warming pushes winter toward more 'comfortable' conditions and the region sees fewer extreme heat days than the national increase; some new crop opportunities emerge with a longer frost-free season."*

Source `https://nca2023.globalchange.gov/key-messages`

Failure DNS resolution failure. The domain hosted the Fifth National Climate Assessment (NCA5, 2023) by the U.S. Global Change Research Program; it was decommissioned in June 2025. The model-generated URL included a `?utm_source=openai` parameter, indicating routing through OpenAI's search infrastructure.

IPA	ASF	SS	CritVM
—	—	—	unevaluable

Verification:
The URL was unreachable at crawl time (DNS resolution failure); all dimensions are unevaluable. Had the source been accessible, the SD5 (Science) × ST1 (Official) cell would yield SS = 5—an optimal citation degraded entirely by an external event (government-site decommissioning). The model generated the URL from training data, but the domain had ceased to exist: a phantom citation driven by infrastructure decay, not by model error.

E DATA RELEASE AND REPRODUCIBILITY

We release CITETRACE as a static snapshot designed to support deterministic re-computation of every metric in Appendix D without re-querying provider APIs. These snapshot-based reproducibility claims refer to the Stack Exchange evaluation. The dataset is hosted on HuggingFace at `https://huggingface.co/datasets/anon-citetrace/CiteTrace`, with collection and analysis code at `https://anonymous.4open.science/r/CiteTrace2027`. This appendix specifies license terms (§E.1), the schema of the released tables (§E.2), and the boundary between what is and is not reproducible from the snapshot (§E.3).

E.1 LICENSING

License terms. Component-specific license terms apply to different components of the release (Table 44). The query set inherits Stack Exchange's CC BY-SA 4.0; research-original outputs are released under CC BY 4.0, and collection and analysis code under the MIT License. Crawling respected robots.txt and per-domain rate limits (§B.3); for redistribution we publish only source URLs and citing-sentence extracts, since per-domain ToS review across 231,105 unique URLs is infeasible.

Table 44: License terms applied to each artifact component. Upstream-derived obligations (CC BY-SA from Stack Exchange) are isolated from research-original outputs (CC BY 4.0) and code (MIT); crawled source contents are withheld.

Artifact	License	Origin / Rationale
Query set (11,200 Stack Exchange titles)	**CC BY-SA 4.0**	Stack Exchange Data Dump
Model response texts & `cited_sentence` extracts	**CC BY 4.0**	Per-provider API ToS
Taxonomy labels (QI/SP/SD/ST/ASF)	**CC BY 4.0**	Research-original output
IPA and SS matrices	**CC BY 4.0**	Research-original output
Aggregate analysis tables (`analysis_master`, etc.)	**CC BY 4.0**	Research-original output
Collection and analysis code	**MIT**	Permissive open-source
Crawled source contents	**Not redistributed**	Per-domain ToS review infeasible

Model-response redistribution. As of the review date (2026-04-24), each of the five providers' (OpenAI, Anthropic, Google, xAI, Perplexity) API Terms of Service grants response ownership to the API caller and does not prohibit academic redistribution of generated content. We verified this provider-by-provider. Reusers must attribute the originating provider and model and comply with each provider's ToS at reuse, which may differ from the terms reviewed on that date.

E.2 SCHEMA AND FIELD DOCUMENTATION

The analysis_master table. The primary release artifact is `analysis_master.parquet`, containing 761,495 rows, one per evaluable citation triple from the final pool. Table 45 documents all 20 columns: each row carries the citation identity (`cit_id`, `query_id`, `model_short`, `url_id`), the citing-sentence content, the five classification labels (`QI_label`, `SP_label`, `ASF_label`, `SD_label`, `ST_label`) with their derived scores (`ipam_score`, `asf_score`, `ssm_score`). This single table is sufficient to reproduce all aggregate metrics in Appendix D at both aggregation units used throughout the paper: citation-level ($n_{\text{cit}} = 761{,}495$, the row-level default) and response-level ($n_{\text{resp}} = 86{,}782$ after excluding zero-citation responses, grouped by `(query_id, model_short)`).

Auxiliary files. Fourteen auxiliary files accompany `analysis_master.parquet` on the HuggingFace release (Table 46). The four parquet views flatten the same citation data into different analysis units (per-query, per-source, citation-only, and pre-extraction response), so secondary analyses can join along any axis without re-deriving keys. Seven human-evaluation tables release the validation studies underlying Appendix C: IPA and SS matrix-cell ratings (`ipam`/`ssm`, $n = 10$ raters

Table 45: Schema of `analysis_master.parquet` (the primary release table, $n = 761{,}495$ rows). Columns are grouped by pipeline stage. Response-level metrics aggregate by `(query_id, model_short)`, and source-level analyses join to `sources.parquet` via `url_id`.

Column	Type	Definition	Allowed values
Row identity			
`cit_id`	INT	Citation unique ID	Unique integer; non-contiguous
Query metadata (§B.1)			
`query_id`	STR	Query identifier	Q00001–Q11200
`site`	STR	Stack Exchange site name	28 official names
`category`	STR	Site's official SE category	6 category labels
Response metadata (§B.2)			
`model_short`	STR	Model short name	10 model labels
`provider`	STR	Model provider	5 provider labels
`cited_sentence`	TEXT	Citing sentence (incl. context)	extracted span
Source metadata (§B.3)			
`url_id`	STR	Source URL unique ID	S0000014–S0435410
`source_url`	STR	Normalized source URL	UTM-stripped, redirect-resolved
`clen`	INT	Crawled body length (chars)	≥ 100
`cited_len`	INT	Citing-sentence length (chars)	≥ 20
`crawl_yn`	STR	Crawl success flag	Y (table includes Y only)
Classification labels (Appendix C)			
`QI_label`	STR	Query Intent label	QI1–QI5
`SP_label`	STR	Source Purpose label	SP1–SP6
`ASF_label`	STR	Answer-Source Fidelity	ASF1–ASF5
`SD_label`	STR	Source Domain label	SD1–SD10
`ST_label`	STR	Source Type label	ST1–ST6
Derived scores (Appendix C)			
`ipam_score`	INT	IPAM[QI][SP] value	1–5
`asf_score`	INT	ASF rubric value	1–5
`ssm_score`	INT	SSM[SD][ST] value	1–5

Table 46: Auxiliary files released alongside `analysis_master.parquet`.

File	Rows	Role
Parquet views		
`queries.parquet`	11,200	Per-query metadata
`sources.parquet`	231,105	Per-source metadata
`citations.parquet`	761,495	Citation-only view of `analysis_master`
`model_responses.parquet`	112,000	Raw responses prior to citation extraction
Human-evaluation tables (Appendix C)		
`ipam_human_eval.parquet`	30	IPA matrix-cell ratings (QI×SP)
`ssm_human_eval.parquet`	60	SS matrix-cell ratings (SD×ST)
`qi_human_eval.parquet`	200	LLM-judge validation, QI task
`sp_human_eval.parquet`	200	LLM-judge validation, SP task
`sd_human_eval.parquet`	200	LLM-judge validation, SD task
`st_human_eval.parquet`	200	LLM-judge validation, ST task
`asf_human_eval.parquet`	200	LLM-judge validation, ASF task
Reference inputs		
`scoring_matrices/ipam_matrix.tsv`	30	QI×SP → IPAM scores (1–5)
`scoring_matrices/ssm_matrix.tsv`	60	SD×ST → SSM scores (1–5)
`site_topology/se_official_audience.json`	28	Stack Exchange site definitions

per cell) and per-task LLM-judge classification ratings (`qi`, `sp`, `sd`, `st`, `asf`; 200 stratified samples × 3 annotators per task). The two scoring matrices (TSV) and the Stack Exchange site definitions (JSON) are reference inputs to the analysis pipeline; per-file schemas are documented in the Croissant manifest (Akhtar et al., 2024; Guha et al., 2016) of the dataset card.

E.3 Reproducibility Notes

Two paths to reproducibility. We provide two distinct paths to reproducibility, since bit-for-bit reproduction of the full pipeline is infeasible: each provider's search tool issues live web queries during generation, so the source set returned for a given query depends on the search-index state at execution time, and running the collection script today against the same 11,200 queries would yield a different response corpus. First, we redistribute the full collected corpus (response texts, citing-sentence extractions, source URLs, taxonomy labels, and matrix scores) so that all metrics in Appendix D can be reproduced deterministically without re-querying provider APIs. Second, the same pipeline (collection scripts, LLM-judge prompts, and the IPA and SS matrices) can be applied to new query sets, new models, or future re-collections, yielding analogous `analysis_master`-format outputs that the released `reproduce.py` can analyze.

Computational requirements. Reproduction runs on a single CPU without a GPU because the released artifact does not re-run model inference. Dependencies are listed in `requirements.txt`.

API usage and token accounting. Response collection issued 112,000 API calls (11,200 queries × 10 models) using integrated search across five providers. Source crawling succeeded for 231,105 of 435,411 URLs. LLM-judge classification issued 1,466,010 calls to `gpt-4o-mini-2024-07-18`: 693,315 source-level calls (231,105 sources × 3 tasks), 11,200 query-level calls, and 761,495 citation-level calls. Estimated usage was 2.7B input and 190M output tokens, excluding prompt development and pilot runs.