

Show Me How You Cite: Citation Fidelity, Usage, and Topology in the Evidence Behavior of AI Scientific Writing

Jihwan Oh^{1,*}, Gyeonghun Sun^{1,*}, Jisoo Kim¹, Yuho Lee¹
Taewon Yun¹, Jisu Shin¹, Hwanjun Song^{1,2,†}

¹Korea Advanced Institute of Science and Technology, ²Cluvion
{jh.oh, skhoon4u, songhwanjun}@kaist.ac.kr

Abstract

Large language models (LLMs) are increasingly used in scientific writing, yet their citation behavior remains poorly understood. We introduce CITESPECTRUM, a controlled dataset for analyzing citations across five levels of AI involvement, from LLM-Only generation to Human-Authoring papers. Using citation-behavior features across fidelity, usage, and topology, we show that AI-generated papers differ from human-written papers in how citations are validated, grounded, used, and structurally organized. Even when retrieval and human guidance reduce ghost citations, AI-generated papers contain more unsupported claims, rely on front-section evidence, use citations more affirmatively, and form broader but less interconnected reference networks. Moreover, citation behavior alone enables a lightweight classifier to outperform strong text-based detectors, achieving 0.880 balanced accuracy and 0.944 AUROC. These results suggest that citation behavior provides a strong, interpretable signal of AI involvement beyond conventional textual patterns. The dataset is available at: <https://github.com/DISL-Lab/CiteSpectrum>.

1 Introduction

Large language models (LLMs) are automating scientific writing rapidly, from editing assistance to generating outlines and subsections (Lee et al., 2022; Wang et al., 2024; Bao et al., 2025), and even full manuscripts (Weng et al., 2025; Lu et al., 2026). A recent audit of ICLR 2026 submissions estimated that 39% contain at least 10% AI-generated content, with 9% being more than 50% AI-generated (Emi, 2025). This rapid integration raises concerns about authorship, attribution, and the reliability of published claims, making it increasingly urgent to characterize how AI- and human-authored scientific writing systematically differ.

Early studies have approached this problem by analyzing intrinsic textual signals of AI-generated writing, such as stylistic features (Jiang et al., 2025; Zhang and Zhou, 2026), rewriting modifications (Mao et al., 2024), and token-level likelihood patterns under language models (Mitchell et al., 2023). These signals have proven effective at exposing the textual traces left by AI generation. However, they focus on the text itself and overlook how AI systems engage with prior scientific knowledge. This engagement is most directly reflected in *citations*, which ground claims in prior work, situate arguments within a field, and reveal how authors reason over scientific evidence.

In this regard, recent work has begun probing how AI systems cite. One line investigates whether AI-generated references are valid, documenting fabricated citations in LLM outputs (Li et al., 2026; Sakai et al., 2026; Xu et al., 2026). Others examine which references AI systems select, finding that LLMs broadly mirror the structure of human citations, while amplifying biases toward highly cited work (Algaba et al., 2025b,a; Mobini et al., 2026). While valuable, these efforts cover only partial aspects of citation behavior, such as reference validity or selection. They reduce AI authorship to a binary, ignoring how citation behavior shifts as AI involvement grows. And they analyze citations only in restricted setups, such as citation-erased prose. As a result, it remains unclear when, how, and along which dimensions AI citation behavior diverges from human practice.

We present the first study of citation behavior across a *controlled spectrum* of AI involvement in scientific writing. As shown in Figure 1, we define this spectrum as five progressive settings, ordered by the amount of human guidance available: (i) LLM-Only, generation from parametric knowledge alone; (ii) RAG, generation with retrieved literature; (iii) Outline-Guided, generation conditioned on a human-curated outline and reference pool; (iv)

*Equal Contribution.

†Corresponding Author.

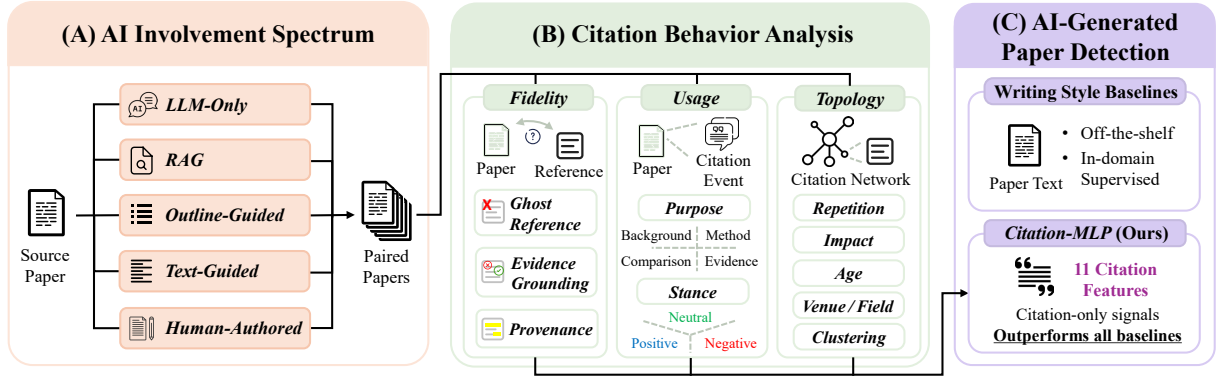


Figure 1: **Overview of our citation-centered study design.** We construct a controlled dataset of paired papers across five levels of AI involvement, from LLM-Only to Human-Authored. For each paper, we analyze citation behavior along three dimensions: *fidelity*, *usage*, and *topology*, yielding 11 citation-behavior feature groups. We then use these features to evaluate whether citation behavior can serve as a signal for AI-generated paper detection.

Text-Guided, generation guided by human prose with the reference pool; and (v) Human-Authored, the original human authorship. This enables us to examine stepwise changes in citation behavior across staged writing workflows, from model-only generation to retrieval, outline-guided writing, text-guided writing, and human authorship.

Beyond detecting whether citation behavior differs, we analyze how it differs along three complementary axes: *Citation Fidelity*, whether cited references really exist and support what is claimed; *Citation Usage*, the rhetorical role each citation plays in the argument; and *Citation Topology*, structural properties of the citation graph, including which references are chosen, how often they are reused, and how they cluster. In total, we introduce 11 citation-behavior feature groups spanning these three dimensions, in contrast to prior work that has examined them in isolation.

Our analysis reveals not only where AI and human citation behavior diverge, but also what kind of guidance closes each gap and to what extent. In citation fidelity, even among verified references, AI leaves a far larger share of claims unsupported, and this gap closes gradually with each added form of guidance, with unsupported-claim rates dropping from 36% in RAG to 14% in Human-Authored. In citation usage, AI cites almost exclusively in support of claims, rarely employing citations for comparison or critique, and this gap closes only when human prose is provided in Text-Guided, while neither retrieval in RAG nor a human-curated reference pool in Outline-Guided is sufficient. In citation topology, AI citation networks are surface-diverse but structurally sparse, repeating references and

clustering less coherently than human networks, and this gap narrows only partially, even under the strongest human guidance.

Taken together, these findings show that AI scientific writing exhibits citation behavior that departs substantially from human evidence use. To empirically validate this citation-level signal, we test whether citation behavior alone can distinguish AI- from human-authored papers on CITESPECTRUM, our controlled dataset of 550 papers (450 AI-generated and 100 human-written) and 78,016 citation events spanning both survey and method papers. A lightweight classifier trained purely on our citation-behavioral features achieves 0.880 balanced accuracy (bAcc) and 0.944 AUROC, outperforming strong text-based detectors, including zero-shot LLM classifier (Qwen3-80B), a fine-tuned DistilBERT detector (Zhang and Zhou, 2026), and the commercial Pangram detector (Emi and Spero, 2024), confirming how AI papers cite carries authorship signals beyond what surface text reveals.

Our main contributions are: (1) we present the first study of citation behavior across a controlled spectrum of AI involvement in scientific writing, moving beyond the binary AI-vs-human framing of prior work; (2) we design 11 citation-behavioral feature groups along three complementary dimensions—fidelity, usage, and topology—and through them reveal substantial and dimension-specific differences between AI- and human-authored scientific writing; and (3) building on these features, we show that citation behavior alone outperforms strong text-based detectors, providing a interpretable and complementary signal for analyzing AI scientific writing.

2 Related Work

AI Writing & Citation Behavior. Recent systems increasingly automate scientific writing, from assisted drafting and survey generation (Lee et al., 2022; Bao et al., 2025) to autonomous manuscript-level pipelines (Weng et al., 2025; Lu et al., 2026). Related work on literature-grounded generation studies citation generation, citation benchmarks, and survey synthesis (Wang et al., 2021b; Funkquist et al., 2023; Wang et al., 2024), but mainly evaluates generation quality rather than how AI involvement changes citation behavior inside complete papers. In parallel, citation analysis has studied rhetorical function and citation stance (Jurgens et al., 2018; Nicholson et al., 2021), reference structure and diversity (Lo et al., 2020; Wang et al., 2021a), and AI-specific citation validity or selection bias (Algaba et al., 2025a,b; Xu et al., 2026). These strands motivate our citation fidelity, usage, and topology dimensions, but are typically studied separately. We instead analyze citation behavior through a controlled spectrum of AI involvement, while consolidating existing citation features into three complementary dimensions.

AI-Generated Paper Detection. AI-generated text detection has largely focused on intrinsic textual traces, including token-likelihood patterns (Mitchell et al., 2023; Bao et al., 2024), rewriting-based signals (Mao et al., 2024; Thai et al., 2026), and supervised stylistic or neural classifiers (Jiang et al., 2025; Zhang and Zhou, 2026). These methods characterize how AI writing looks, but not how it engages with scientific evidence. Recent studies use citation patterns to distinguish AI-generated from human-written papers or analyze structural anomalies in AI reference networks (Algaba et al., 2025b; Mobini et al., 2026). However, their primary focus is detecting AI authorship or identifying aggregate citation anomalies, rather than explaining how citation behavior changes under different forms of AI involvement. Our work uses detection as an empirical validation, while focusing on how citation fidelity, usage, and topology shift from retrieval-based generation to human-guided writing and human authorship.

3 Study Design for Citation Analysis

Figure 1 illustrates our study design. From a fixed source topic, we construct a controlled five-level AI-involvement spectrum (Section 3.2) and an-

alyze the resulting papers along three citation-behavior dimensions: *fidelity*, *usage*, and *topology* (Section 3.3). This design reveals which forms of guidance recover human-like citation practice and which behaviors remain difficult to reproduce.

3.1 Scientific Corpus Construction

Since existing studies typically compare AI-generated and human-written papers as two endpoints, it remains difficult to isolate which stages of the writing process drive changes in citation behavior. Moreover, prior citation analyses often focus on individual dimensions, such as citation validity, rhetorical intent, or reference graph structure, without a controlled setting that traces how these behaviors evolve as human guidance increases.

To enable fine-grained analysis, we construct CITESPECTRUM, a controlled dataset of 550 scientific papers with 78,016 citation events, to characterize citation behavior across five levels of AI involvement using 11 feature groups.

CITESPECTRUM consists of two subsets: (i) survey papers, which enable fine-grained citation behavior analysis across the full AI involvement spectrum, and (ii) method papers, which extend the analysis to a broader paper type under the binary AI/Human setting. Both subsets are used for citation behavior analysis and AI/Human detection. Dataset statistics are provided in Appendix C.1.

Survey Paper Generation. The survey data consist of 50 human-written surveys sampled across 14 domains in SurveyGen (Bao et al., 2025), including computer science, medicine, biology, and psychology. Each survey serves as a source paper from which we derive the components needed to construct the five AI-involvement conditions: a fixed topic, outline, prose, and bibliography. These components are selectively provided to the model at each level, enabling controlled variation in writing guidance (see Section 3.2). For each survey, we use two LLMs, GPT-5.4 and Qwen3-Next-80B-A3B-Instruct, to generate AI-involved variants, ensuring that observed patterns are not specific to a single model family (detailed configurations are provided in Appendix E.1). Because all conditions are generated from the same source survey, topic and domain are held fixed across the spectrum, making citation-behavior differences more attributable to the level of AI involvement.

We center our main analysis on the survey subset because surveys are literature-centered: retrieving,

Component	LLM	RAG	Outline	Text	Human
Refs.	Model	Retrieved	Guided	Guided	Human
Outline	Model	Model	Guided	Guided	Human
Prose	Model	Model	Model	Guided	Human
Citation	Model	Model	Model	Model	Human

Table 1: Human guidance across writing conditions. *Model*: stage produced by the model without human input; *Retrieved*: externally retrieved literature; *Guided*: model generation conditioned on a human artifact; *Human*: direct human authorship.

organizing, synthesizing, and citing prior work constitutes their core task, making citation behavior a direct rather than peripheral signal.

Method Paper Collection. The method subset is composed of 100 papers collected from existing sources rather than generated: 50 human-written method papers from ICLR 2026 oral presentations and 50 AI-generated method papers from ICAIS 2025, AI Scientist (Lu et al., 2026), and FARS (Tang et al., 2026). Unlike surveys, method papers involve additional factors such as idea generation, implementation, experimental execution, and result validation, which make fine-grained control over AI involvement substantially harder. We therefore use the method subset only in the binary AI/Human setting, treating it as a complementary testbed for examining whether citation-behavior patterns observed in surveys generalize to a broader paper type.

3.2 Human Guidance in AI Writing

Starting from each source survey, we generate five paper variants that span a spectrum of human guidance. Moving from LLM-Only to Human-Authored, each successive condition supplies progressively more components of the source survey (topic, outline, prose, and bibliography) as guidance, while the remaining stages stay model-driven. Table 1 summarizes how each writing stage is produced under each condition.

LLM-Only. The model receives only the source survey topic, without an external reference list, human outline, or human prose. It first generates a survey outline from the topic, then writes the survey section by section. During writing, the model must independently decide which scholarly papers to cite, generate the corresponding reference entries, and place in-text citation markers. This model-only condition produces a full survey from parametric knowledge and is used primarily to measure ghost

references under unguided generation. Implementation details are in Appendix E.2.

RAG. Starting from LLM-Only, this condition adds a retrieved reference pool constructed from external scholarly search.¹ The model receives the source topic and retrieved references, but not the human outline, human bibliography, or human prose. It generates an outline conditioned on the retrieved literature, selects subsection-level references, and writes each subsection with citations linked to the selected references. Thus, external literature access is added, while organization, prose generation, within-pool reference selection, and citation placement remain model-driven. Implementation details are in Appendix E.3.

Outline-Guided. Compared with RAG, this condition replaces retrieval-driven reference discovery and model-generated organization with human guidance. The model receives the source topic, a normalized outline derived from the human survey, and a human-selected reference pool mapped to sections or subsections, but not the original human prose. These artifacts provide human structure and locally mapped references. For each subsection, the model first selects locally relevant references from the assigned human-curated pool using titles and abstracts; the full texts of the selected references are then provided for subsection writing under the human-derived outline. This condition tests whether human organization and reference selection close citation gaps that retrieval alone cannot close. Implementation details are in Appendix E.4.

Text-Guided. Building on Outline-Guided, this condition further provides human-written survey prose with citation markers withheld. For each subsection, the model first selects relevant references from the provided human-curated pool using titles and abstracts; the full texts of the selected references are then provided while it rewrites the given prose and restores citation markers, preserving the original discourse and argument flow as much as possible. This condition tests whether human-written discourse closes citation gaps that remain after human organization and reference selection. Implementation details are in Appendix E.5.

Human-Authored. This condition is the original human-written survey. The topic, organization, reference selection, prose, and citation placement are

¹We use the Semantic Scholar API for retrieval; the model selects subsection-level references before writing.

all determined by human authors. It serves as the endpoint of the spectrum and the reference point for human citation behavior.

Taken together, the five conditions compare citation behavior across staged writing workflows with progressively increasing human guidance. This lets us ask which citation gaps are reduced by external literature access, which require human-curated structure and references, and which persist until the model is given human-written discourse.

3.3 Citation-Behavior Features

We instantiate *citation fidelity*, *usage*, and *topology* with 11 paper-level features. Unlike prior work that examines these dimensions in isolation, we jointly capture all three, enabling factual, rhetorical, and structural views of citation behavior within the same study. All features are interpretable, stable at the paper level, and applicable to both AI-involved and human-written papers.

Detailed metric definitions and human validation of the LLM-judged fidelity and usage labels are provided in Appendices A–B.

Citation Fidelity measures whether a paper’s citations are bibliographically valid and evidentially grounded. Following prior work on hallucinated or invalid references in LLM outputs (Xu et al., 2026; Sakai et al., 2026; Li et al., 2026) and citation-fidelity verification (Chen et al., 2025; Sadeghi et al., 2026), we use three features:

- *Ghost Reference Rate*: The fraction of unique references that cannot be resolved to real scholarly works. In LLM-Only, this captures references hallucinated from parametric knowledge; in reference-conditioned settings, it captures failures to cite resolvable papers from the provided pool, the set of candidate references.
- *Evidence Grounding Rate*: Whether the cited reference supports the claim made in the citation sentence. For each claim–reference pair, we retrieve relevant evidence paragraphs from the cited paper and judge whether the evidence supports the claim, labeling the pair as supported or unsupported, via an LLM-as-a-judge.
- *Evidence Provenance Distribution*: Where the supporting evidence is located within the cited paper, categorized as introductory, body, or result-oriented sections. This captures whether citations rely on surface-level passages or on deeper experimental or analytical content.

Citation Usage measures the rhetorical role of a citation in the citing text. Following prior work on citation function (Teufel et al., 2006; Jurgens et al., 2018), citation intent (Cohan et al., 2019), and citation stance (Nicholson et al., 2021), we use two event-level label distributions:

- *Purpose Distribution*: Whether a citation supports a specific claim (evidence), provides general context (background), adopts a method, model, or dataset from the cited work (method), or contrasts approaches or findings (comparison), classified via an LLM-as-a-judge.
- *Stance Distribution*: Whether the citing text endorses the cited work (positive), references it without a clear stance (neutral), or challenges, contrasts, or disagrees with it (negative), being classified via an LLM-as-a-judge.

These features capture differences fidelity alone cannot: a citation can be valid and grounded yet used differently, such as affirmative evidence in AI-generated surveys versus contextual or neutral positioning in human-written ones.

Citation Topology measures how a paper selects, distributes, and connects its references at the document level. Following prior work on citation diversity (Wang et al., 2021a), bibliometric citation structure (Lo et al., 2020), and AI reference-selection bias or citation-network anomalies (Mobini et al., 2026), we define topology features as paper-level aggregates, using normalized ratios to reduce the influence of citation volume:

- *Reference Age Gap*: The median publication-year difference between citing and cited papers. Higher values indicate reliance on older works.
- *Venue Diversity Ratio*: The fraction of unique venues among cited works. Higher values indicate broader venue coverage.
- *Field Diversity Ratio*: The fraction of unique research subfields among cited works. Higher values indicate broader topical coverage.
- *Citation Count*: The median citation count of cited works, capturing their prominence. Higher values indicate reliance on highly cited papers.
- *In-text Repetition Rate*: The degree to which a small subset of references accounts for most citation events. Higher values indicate more concentrated reuse.

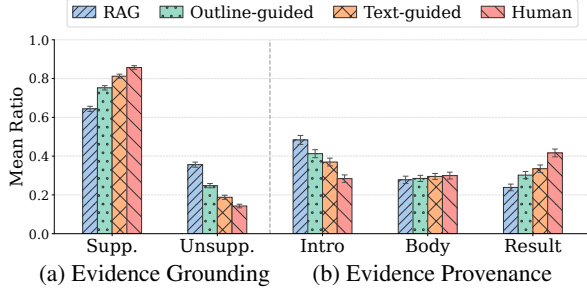


Figure 2: Citation fidelity features across AI involvement levels. Error bars indicate standard errors.

	LLM	RAG	Outline	Text	Human
Ghost (%)	79.2	0.4	0.9	0.6	0.3

Table 2: Ghost reference rate across AI involvement levels. LLM, Outline, Text, and Human denote LLM-Only, Outline-Guided, Text-Guided, and Human-Authored, respectively.

- *Network Clustering Coefficient*: The average clustering of cited references in their citation graph. Higher values indicate a more cohesive, interconnected reference set.

Together, these features distinguish a broad but weakly connected reference portfolio from a more focused and cohesive scholarly context.

4 Analysis: AI vs. Human Citation

We analyze how AI-generated and human-written papers differ in citation behavior, and how these gaps change as human guidance is added. The results show that each added form of guidance reduces some gaps, but AI-generated writing still differs from human writing in most dimensions.

4.1 Patterns on Citation Fidelity

Table 2 shows that LLM-Only suffers from ghost references, with 79.2% of its references failing to resolve to real scholarly works. In contrast, providing external documents largely eliminates this failure: the ghost reference rate drops to 0.4% in RAG, 0.9% in Outline-Guided, and 0.6% in Text-Guided, all close to the 0.3% rate observed in Human-Authored. This suggests that parametric knowledge alone is insufficient for citing real references without external reference guidance. The few ghost references in human-written papers were manually inspected and mostly corresponded to unpublished or unsearchable papers. We exclude LLM-Only from subsequent analyses because its high ghost reference rate makes further citation behavior analysis unreliable.

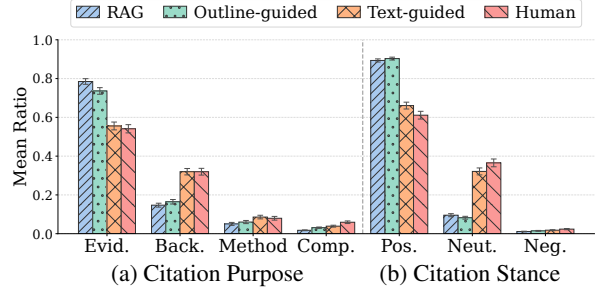


Figure 3: Citation usage features across AI involvement levels. Error bars indicate standard errors.

Figure 2 shows that resolving ghost references does not guarantee faithful claim grounding. In Figure 2(a), only 64% of citation instances in RAG are supported by the cited references, leaving 36% unsupported. Evidence grounding linearly improves as stronger guidance is added: supported citations increase from 64% in RAG to 75% in Outline-Guided, 81% in Text-Guided, and 86% in Human-Authored. This suggests that retrieval can supply real references, but structural and textual guidance are essential to connect those references to claims more accurately.

Figure 2(b) shows that RAG relies most heavily on introductory evidence with 48%, whereas this linearly decreases to 41% in Outline-Guided, 37% in Text-Guided, and 28% in Human-Authored. This suggests that RAG tends to draw on surface-level passages, such as abstracts and introductions. As more specific guidance is provided, models increasingly move beyond these front sections: evidence from results sections increases from 24% in RAG to 42% in Human-Authored. This indicates that structural and textual guidance enables models to ground claims in more specific evidence, such as experimental results and analyses.

4.2 Patterns on Citation Usage

Figure 3 shows that citation usage changes as human guidance increases. In the citation purpose analysis (Figure 3(a)), citations serving as direct evidence for claims dominate in RAG (78%) and Outline-Guided (74%). This suggests that retrieval-based generation tends to use citations primarily as direct evidence for claims, and that adding an outline alone does not substantially change this pattern. In contrast, citations used as direct evidence drop to 56% in Text-Guided and 54% in Human-Authored, while background citations increase to 32% in both settings. This shift suggests that textual guidance makes citation usage more human-like, using citations to contextualize

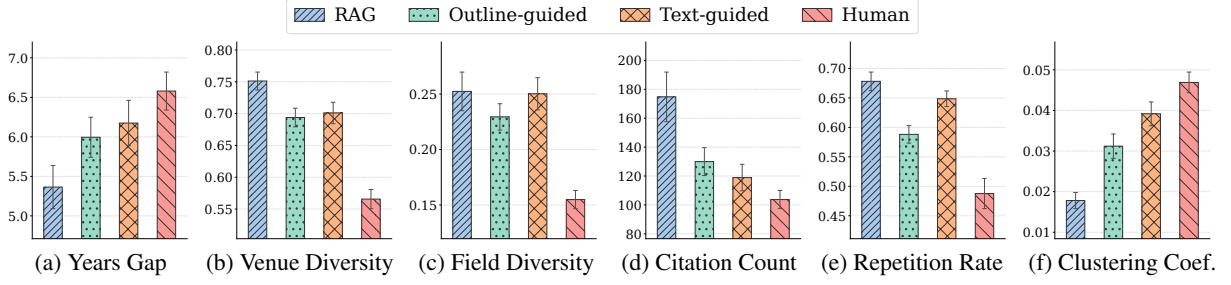


Figure 4: Citation topology features across AI involvement levels. Error bars indicate standard errors.

	Fidelity					Usage								Topology					
	Supp.	Unsupp.	Intro	Body	Result	Back.	Method	Comp.	Evid.	Pos.	Neut.	Neg.		Years	Venue	Field	Cited	Repeat	Cluster
Survey	↓	↑	↑	—	↓	↓	↓	↓	↑	↑	↓	↓		↓	↑	↑	↑	↑	↓
Method	↓	↑	↑	—	↓	↓	↓	↓	↑	↑	↓	↓		↓	↑	↑	↑	↑	↓
Match	✓	✓	✓	—	✓	✗	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✗	✓

Table 3: Survey–method consistency of statistically supported AI–Human citation-behavior differences. Arrows indicate whether AI-generated papers have significantly higher (↑) or lower (↓) values than human-written papers ($p < 0.05$); “—” indicates a non-significant difference. Survey trends are tested using source-level linear mixed-effects models, and method-paper differences using two-sided Welch’s t -tests.

and situate claims within prior work.

In the citation stance analysis (Figure 3(b)), the stance shifts with the level of AI involvement. RAG and Outline-Guided exhibit a strongly positive stance at nearly 90% and neutral stance below 10%. This suggests a support-seeking tendency in retrieval-based writing, where selected documents are framed mainly as confirming evidence rather than neutral context or critical contrast. As human involvement increases, this pattern weakens: Text-Guided reduces positive citations to 66% while increasing neutral citations to 32%, closely approaching Human-Authored.

4.3 Patterns on Citation Topology

Figure 4 shows that citation topology also changes as human guidance increases. The source-profile features in Figure 4(a)–(c) show two different patterns. Reference age changes linearly: the years gap increases from RAG (5.3 years) to Human-Authored (6.5 years) in Figure 4(a), suggesting that human writing draws more on older literature. In contrast, venue and field diversity in Figure 4(b) and Figure 4(c) do not simply move linearly with guidance; instead, they mainly separate human writing from AI-generated settings, suggesting that even textual guidance does not fully reproduce the more focused source selection of human writers.

The remaining topology features in Figure 4(d)–(f) describe how references are connected. Citation count decreases progressively from RAG (175) to Human-Authored (104) in Figure 4(d), suggesting that AI writing tends to prefer highly cited

references. Repetition rate in Figure 4(e) again separates human writing from AI-generated settings: AI-generated writing repeatedly reuses selected references more often, even with textual guidance. Finally, clustering coefficient increases toward Human-Authored in Figure 4(f), showing that human-written papers cite references that are more mutually interconnected. This reflects deeper engagement with a coherent body of literature: human authors draw from a focused research community where cited works build on each other, whereas AI retrieves topically relevant but loosely connected references.

4.4 Generalization across Paper Types

Method papers are constructed only in the binary AI/Human setting because controlled and fine-grained AI-involvement construction is difficult for work involving ideation, experiment design, execution, and result validation (Weng et al., 2025; Lu et al., 2026). To test whether our citation-behavior taxonomy transfers across paper types, we compare AI–Human differences in citation-behavior features for both survey and method papers.

For survey papers, we test ordered trends across RAG, Outline-Guided, Text-Guided, and Human-Authored using source-level linear mixed-effects models, treating the source survey as a random effect. For method papers, which are independently sampled, we compare AI-generated and human-written papers using two-sided Welch’s t -tests. Table 3 reports the directions of statistically significant differences ($p < 0.05$).

Among the 17 subfeatures with statistically significant differences, 15 show consistent directions across survey and method papers. AI-generated papers generally show lower citation support, stronger reliance on front-section evidence, more affirmative citation stance, broader and more highly cited reference selection, and lower reference clustering than human-written papers. Body provenance is not significant in either analysis, while background citation ratio and repetition rate show opposite directions across paper types, suggesting that these behaviors are more sensitive to paper type.

Overall, the strong directional consistency across statistically supported features indicates that the citation-behavior differences observed in Section 4 largely generalize beyond surveys to method papers. We also provide the box plot distributions of all citation features across survey and method generation settings in Figure 6–8.

5 Citation-Based AI Paper Detection

We evaluate whether citation-behavior features can detect AI-involved papers in a binary AI/Human setting. The AI class consists of documents generated under the RAG, Outline-Guided, and Text-Guided conditions, while the Human class corresponds to Human-Authored papers.

Dataset. We use CITESPECTRUM for binary detection. For surveys, the AI class consists of three conditions, RAG, Outline-Guided, and Text-Guided, yielding 300 AI-generated surveys, while the Human class consists of 50 Human-Authored surveys. For method papers, we use the 50 AI-generated and 50 human-written papers from Section 3.1. In total, the dataset contains 350 AI and 100 Human papers. The detection dataset details are provided in Appendix D.1.

Training and Evaluation. We train a lightweight two-layer MLP classifier, CITATION-MLP, using 16 paper-level citation-behavior features instantiated from the 11 feature groups spanning fidelity, usage, and topology. The classifier consists of two fully connected hidden layers with ReLU activation, followed by a sigmoid output for binary AI/Human prediction. We evaluate performance using domain-balanced 5-fold cross-validation and report Balanced Accuracy (bAcc), Macro-F1, and AUROC. We detail the Citation-MLP model configuration in Appendix D.2, and the evaluation protocol in Appendix D.3.

Method	bAcc.	Macro-F1	AUROC
<i>Off-the-Shelf</i>			
Qwen3-80B-Inst.	0.626	0.506	0.650
Fast-DetectGPT	0.611	0.542	0.674
EditLens	0.660	0.471	0.787
Pangram	0.844	0.731	0.816
<i>In-domain Supervised</i>			
DistilBERT	0.775	0.763	0.882
Llama-3.2-3B	0.793	0.792	0.924
Citation-MLP (Ours)	0.880	0.876	0.944

Table 4: Binary AI/Human detection results. Our citation-based classifier is compared against off-the-shelf detectors and in-domain supervised classifiers.

Baseline Methods. We compare CITATION-MLP with six strong text-based baselines spanning off-the-shelf AI detectors and in-domain supervised classifiers. The off-the-shelf group includes Qwen3-Next-80B-A3B-Instruct, a pretrained LLM; Fast-DetectGPT (Bao et al., 2024), an efficient zero-shot detector that uses conditional probability to distinguish machine-generated text; EditLens (Thai et al., 2026), a state-of-the-art open source AI text detector; and Pangram (Emi and Spero, 2024), the strongest commercial detector trained from 28M documents. The supervised group includes DistilBERT (Zhang and Zhou, 2026), a compact transformer encoder fine-tuned as an AI text classifier, and Llama-3.2-3B, representing a LoRA (Hu et al., 2022) fine-tuned LLM classifier. Both supervised baselines are trained on CITESPECTRUM. Configurations are detailed in Appendix D.4.

Detection Performance. As summarized in Table 4, CITATION-MLP achieves the best performance across all metrics, with 0.880 balanced accuracy, 0.876 Macro-F1, and 0.944 AUROC. It improves over the strongest off-the-shelf detector, Pangram, by 3.6 percentage points in balanced accuracy, 14.5 points in Macro-F1, and 12.8 points in AUROC. Other open-source text-based baselines such as Qwen3-80B, Fast-DetectGPT, and EditLens perform substantially worse (bAcc. < 0.7), suggesting limited effectiveness in our controlled scientific-writing setting. It also outperforms in-domain supervised text classifiers such as DistilBERT and Llama-3.2-3B, despite using only citation-derived features. Across all supervised models, performance remains stable across the five folds, with standard deviations at most 0.053 across all metrics. These results indicate that citation behavior provides a strong and interpretable authorship signal beyond conventional text-based cues.

Feature Set	bAcc.	Macro-F1	AUROC
All citation features	0.880	0.876	0.944
w/o Fidelity	0.801	0.811	0.895
w/o Usage	0.801	0.797	0.897
w/o Topology	0.748	0.748	0.856
Fidelity only	0.660	0.656	0.760
Usage only	0.705	0.701	0.769
Topology only	0.743	0.743	0.850

Table 5: Feature contribution analysis for CITATION-MLP. Each row uses the same classifier and evaluation protocol, varying only the input feature group.

Feature Contribution Analysis. We examine the contribution of each dimension in our citation taxonomy through feature-group ablations. Following Section 3.3, we group the 11 features and 18 subfeatures into three dimensions: *fidelity*, *usage*, and *topology*. Using the same CITATION-MLP and field-balanced 5-fold cross-validation, we evaluate the full feature set, remove each dimension in turn, and evaluate each dimension individually.

As shown in Table 5, the full feature set achieves the best performance across all metrics. Removing topology causes the largest degradation, reducing balanced accuracy from 0.880 to 0.748 (13.2 percentage points), Macro-F1 from 0.876 to 0.748 (12.8 points), and AUROC from 0.944 to 0.856 (8.8 points). Removing fidelity or usage also lowers balanced accuracy by 7.9 percentage points, indicating that both provide complementary signals. Among the individual dimensions, topology performs best, reaching both 0.743 balanced accuracy and Macro-F1, whereas fidelity provides the weakest standalone signal, with 0.660 balanced accuracy and 0.656 Macro-F1.

6 Conclusion

We analyze citation behavior across a spectrum of AI involvement in scientific writing, ranging from fully LLM-generated to fully human-written papers, and find that AI-generated papers diverge systematically from human-written ones in citation fidelity, usage, and topology. Specifically, AI-generated papers exhibit more unsupported claims, greater reliance on front-section evidence, more affirmative framing, and broader but less interconnected references. We further show that these citation-level distinctions enable accurate detection of AI involvement in scientific writing, establishing citation behavior as a strong, interpretable signal of AI authorship beyond conventional textual patterns.

Limitations

Scope of the AI-Involvement Spectrum. Our fine-grained AI-involvement spectrum focuses on survey papers because it is fundamentally literature-centered, with retrieval, organization, synthesis, and drafting forming a relatively staged workflow in which writing guidance can be controlled directly. In contrast, method research involves broader and inherently iterative sources of AI involvement, including problem formulation, method design, implementation, experimentation, and result analysis. For example, implementation failures or experimental outcomes may trigger revisions to the method, setup, or even the research question. Consequently, imposing the same ordinal spectrum on method papers would confound writing-stage guidance with evolving research-process variables, while citation behavior captures only one dimension of this broader process. We therefore use a binary AI/Human setting for method papers and test whether survey-derived citation patterns transfer to this genre (Section 4.4) and effective AI paper detection (Section 5).

Human-Authored Papers as a Reference Class.

Human-authored papers are an empirical reference class, not a normative gold standard: human citation practice can also be biased, shallow, weakly grounded, or partially shaped by AI assistance. Thus, observed AI–Human differences should be interpreted as behavioral distinctions rather than evidence that human citation practice is inherently superior. Moreover, as AI-assisted scientific writing becomes more prevalent, verifying the absence of AI involvement in human-authored papers may become increasingly difficult, complicating the construction of a strictly human-only reference class.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-NR068758). Additionally, this work was supported by the Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Korea government (MSIT) (No. RS-2026-25492133); and by the Korea Institute of Science and Technology Information (KISTI) in 2026 (No. (KISTI)K26L3M1C1); and InnoCORE program of the Ministry of Science and ICT (No. N10260110, AI MetaScientist).

References

- Andres Algaba, Vincent Holst, Floriano Tori, Melika Mobini, Brecht Verbeken, Sylvia Wenmackers, and Vincent Ginis. 2025a. How deep do large language models internalize scientific literature and citation practices? *arXiv preprint arXiv:2504.02767*.
- Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025b. Large language models reflect human citation patterns with a heightened citation bias. In *NAACL Findings*.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *ICLR*.
- Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. Surveygen: Quality-aware scientific survey generation with large language models. In *EMNLP*.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP*.
- Hong Chen, Misha Teplitskiy, and David Jurgens. 2025. The noisy path from source to citation: Measuring how scholars engage with past research. In *ACL*.
- Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL*.
- Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*.
- Bradley Emi. 2025. [Pangram predicts 21% of ICLR reviews are AI-generated](#). Pangram Labs blog post.
- Bradley Emi and Max Spero. 2024. Technical report on the pangram AI-generated text classifier. *arXiv preprint arXiv:2402.14873*.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2023. CiteBench: A benchmark for scientific citation text generation. In *EMNLP*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*.
- Lei Jiang, Desheng Wu, and Xiaolong Zheng. 2025. SenDetEX: Sentence-level AI-generated text detection for human-AI hybrid content via style and context fusion. In *EMNLP*.
- David Jurgens, Srikanth Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.
- Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, and 29 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.
- Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *CHI*.
- Peiran Li, Fangzhou Lin, Shuo Xing, Xiang Zheng, Xi Hong, Siyuan Yang, Jiashuo Sun, Zhengzhong Tu, and Chaoqun Ni. 2026. BibAgent: An agentic framework for traceable miscitation detection in scientific literature. *arXiv preprint arXiv:2601.16993*.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *ACL*.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919.
- Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: geneRative AI detection via rewriting. In *ICLR*.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *ICML*.
- Melika Mobini, Vincent Holst, Floriano Tori, Andres Algaba, and Vincent Ginis. 2026. Structurally human, semantically biased: Detecting llm-generated references with embeddings and gnns. In *ICLR*.
- Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. 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.
- Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

- Shaghayegh Sadeghi, Khashayar Khajavi, Rise Adhikari, and Alexander Tessier. 2026. Deepsciverify: Verifying scientific claim–citation alignment via llm-driven evidence escalation. *arXiv preprint arXiv:2605.27710*.
- Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Hallucination matters: Revealing the impact of hallucinated references with 300 hallucinated papers in acl conferences. *arXiv preprint arXiv:2601.18724*.
- Qiong Tang, Tianxiang Sun, Xiangkun Hu, Xiangyang Liu, Yiran Chen, Yunfan Shao, Bobo Li, Changze Lv, Cheng Xu, Chengsong Huang, Chunyang Li, Dizhan Xue, Hao Bai, Haodong Duan, Hengquan Guo, Hongyang He, Hongyi Chen, Hui Shen, Jiahao Yuan, and 38 others. 2026. Fars: A fully automated research system deployed at scale. *arXiv preprint arXiv:2606.31651*.
- Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *EMNLP*.
- Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. 2026. Editlens: Quantifying the extent of AI editing in text. In *ICLR*.
- Lucy Lu Wang, Kelly Mack, Emma J McDonnell, Dhruv Jain, Leah Findlater, and Jon E. Froehlich. 2021a. A bibliometric analysis of citation diversity in accessibility and hci research. In *CHI*.
- Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. 2021b. Autocite: Multi-modal representation fusion for contextual citation generation. In *WSDM*.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: Large language models can automatically write surveys. In *NeurIPS*.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *ICLR*.
- Zuyao Xu, Yuqi Qiu, Lu Sun, FaSheng Miao, Fubin Wu, Xiang Li, Xinyi Wang, Haozhe Lu, ZhengZe Zhang, Yuxin Hu, Jialu Li, Luo Jin, Feng Zhang, Rui Luo, Xinran Liu, Yingxian Li, and Jiaji Liu. 2026. Ghostcite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718*.
- Alfredo Yegros-Yegros, Ismael Ràfols, and Pablo D’Este. 2015. Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PLOS ONE*.
- Cuilian Zhang and Weijun Zhou. 2026. Efficient detection of AI-generated scientific abstracts with a lightweight transformer. *Scientific Reports*, 16(4975).

A Human Evaluation Details

We conduct a human evaluation to validate the citation labels used in our citation-behavior taxonomy. We evaluate three citation-event labels: *evidence grounding* in *fidelity*, which captures whether the cited reference supports the citation-backed claim; *purpose*, which captures why a citation is introduced; and *stance* in *usage*, which captures the citing author’s attitude toward the cited work.

LLM Usage Labeling. Before human validation, citation usage labels are generated with GPT-5.4-mini. For each citation event, the model receives the section title, coarse section bucket, bundle indicator, target citation marker group, citing sentence, surrounding citation block, and linked reference titles. The target citation marker group and citing sentence are explicitly included in the prompt, so that when multiple citation markers appear in the same sentence, the model labels only the event corresponding to the target marker group. The model returns both *purpose* and *stance* labels in strict JSON format. The full labeling prompt is shown in Table 17.

Annotation Sample. We use the same annotation sample for all three tasks. Specifically, we sample 1,000 citation events from four writing conditions: RAG, Outline-Guided, Text-Guided, and Human-Authored. For each condition, we sample 25 survey papers and extract 10 citation events per paper, resulting in 250 citation events per condition and 1,000 citation events in total.

Annotation Procedure. For *evidence grounding*, annotators are shown a cited reference and determine whether the reference supports the claim. The possible labels are supported and unsupported as binary. Because this task requires academic reasoning over the cited reference, it is annotated by three in-house PhD students.

For each citation event, annotators are shown the citation sentence and its local context, including the previous and next sentence when available. The target citation marker is highlighted in the citation sentence. For *purpose*, annotators assign one of four labels: background, method, comparison, and evidence. For *stance*, annotators assign one of three labels: positive, neutral, and negative. These are annotated by three Amazon Mechanical Turk workers per citation event, with a total crowdsourcing cost of \$30. We recruited Amazon Mechanical Turk workers residing in Australia (AU), Canada

Metric	Grounding	Purpose	Stance
<i>Human Agreement</i>			
Pairwise agreement	0.964	0.897	0.919
Fleiss’ κ	0.828	0.701	0.701
<i>Automated Validation</i>			
Accuracy	0.971	0.971	0.977
Macro-F1	0.957	0.941	0.890
Cohen’s κ	0.914	0.895	0.893

Table 6: Validation of citation labels, reporting human-human agreement and automated-human agreement.

(CA), New Zealand (NZ), Great Britain (GB), and the United States (US). To ensure annotation quality, we restricted participation to MTurk workers with an approval rate above 95%.

Metrics. To assess annotation reliability, we report average pairwise annotator agreement and Fleiss’ κ (Fleiss, 1971) for each task.

Results. Table 6 shows high agreement across all three citation-labeling tasks. For human annotation, pairwise agreement is 0.964 for grounding, 0.897 for purpose, and 0.919 for stance, with Fleiss’ κ values of 0.828, 0.701, and 0.701, respectively. The higher agreement for grounding should be interpreted in light of its binary label space, compared with the multi-class purpose and stance tasks. Automated labels also show strong agreement with adjudicated human labels. Grounding achieves an accuracy/Macro-F1 of 0.971/0.957, purpose 0.971/0.941, and stance 0.977/0.890, with Cohen’s κ values ranging from 0.893 to 0.914. These results support the reliability of the automated labels used in our main analysis.

B Taxonomy Details

B.1 Citation Fidelity Metrics

We compute citation fidelity using three metrics: ghost reference rate, evidence grounding rate, and evidence provenance distribution. These metrics measure whether cited references are real scholarly works, whether citation-backed claims are supported by the cited references, and where the supporting evidence appears within the cited papers.

Ghost Reference Rate. We first verify whether each parsed reference corresponds to a real scholarly work. Reference resolution is performed in four stages. Given the parsed reference title, we first search Google Scholar², Semantic Scholar (Kinney

²<https://scholar.google.com>

et al., 2023), and OpenAlex (Priem et al., 2022) using each API. If the reference is not resolved by these three sources, we perform an additional web search using GPT-5.4-mini³ with the full parsed reference metadata, including title, authors, venue, and year when available. References that cannot be resolved after all four stages are labeled as ghost references, while all resolved references are treated as real references. The Ghost Reference Rate is defined as:

$$\text{Ghost Reference Rate} = \frac{|\mathcal{R}_{\text{ghost}}|}{|\mathcal{R}|},$$

where $\mathcal{R}$ is the set of all references in a paper and $\mathcal{R}_{\text{ghost}}$ is the subset of references that cannot be resolved to real scholarly works.

Evidence Grounding Rate. We construct claim-reference pairs from citation sentences and their associated cited references. For each claim-reference pair, we retrieve the top-5 most relevant paragraphs from the cited reference using bge-m3⁴ embedding model. A claim is considered supported only when at least one of the retrieved paragraphs supports the claim. We determine support using two complementary criteria: a GPT-5.4-mini support judgment based on the prompt in Table 16, and an mDeBERTa⁵ NLI model judgment based on whether at least one of the top-5 paragraphs entails the claim. We label a claim-reference pair as supported only when both criteria agree that the reference supports the claim; otherwise, it is labeled as unsupported. The Evidence Grounding Rate is defined as:

$$\text{Evidence Grounding Rate} = \frac{|\mathcal{P}_{\text{sup}}|}{|\mathcal{P}|},$$

where $\mathcal{P}$ is the set of all claim-reference pairs and $\mathcal{P}_{\text{sup}}$ is the subset of pairs labeled as supported.

Sensitivity to Evidence Retrieval. To assess whether top-5 paragraph retrieval misses relevant evidence, we repeat the grounding evaluation using the full cited paper as input. Full-text evaluation increases support rates by only 2–6 percentage points across conditions, while preserving the ordering of all survey and method conditions. Under full-text evaluation, adjacent survey conditions remain separated by at least 7 percentage points, and the AI–Human gap for method papers remains 11 points,

³<https://openai.com>

⁴<https://huggingface.co/BAAI/bge-m3>

⁵<https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>

suggesting that the top-5 retrieval choice does not drive our main grounding results.

Evidence Provenance Distribution. Evidence Provenance Distribution measures where supported evidence is located within the cited reference. We first parse each reference paper using OCR and extract its section-level text and header names. We then map each section header into one of three coarse provenance categories: *introduction*, *body*, or *result*, using SciBERT (Beltagy et al., 2019). For each supported claim-reference pair, we identify the retrieved paragraph that provides support and assign its section category as the provenance label. The provenance distribution is computed over supported claim-reference pairs:

$$\text{Prov}_c = \frac{|\{p \in \mathcal{P}_{\text{sup}} : \text{section}(p) = c\}|}{|\mathcal{P}_{\text{sup}}|},$$

where $c \in \{\text{Intro}, \text{Body}, \text{Result}\}$.

This yields the proportion of supported citations whose evidence comes from introductory, body, or result sections of the cited references.

B.2 Citation Usage Metrics

This section provides implementation details for the citation usage metrics introduced in Section 3.3. Our usage labels build on prior work on citation function and rhetorical role (Teufel et al., 2006; Jurgens et al., 2018; Cohan et al., 2019) and citation stance (Nicholson et al., 2021).

We define a citation event as a citation marker cluster within a sentence. Multiple references grouped in the same marker cluster are treated as one event, while multiple marker positions in the same sentence are split into separate events. For each event, we extract the citing sentence and its neighboring context, consisting of the previous, current, and next sentence when available. The linked reference titles and the marker group are also provided to the labeler.

Purpose Distribution. Each citation event is assigned one of four purpose labels. *Background* indicates that the citation provides general context or prior knowledge. *Method* indicates that the citing work directly uses or adopts a method, model, dataset, software, or concept from the cited work. *Comparison* indicates that the citation is used to compare or contrast approaches, findings, or systems. *Evidence* indicates that the citation supports or justifies a specific claim or interpretation. For

each paper, we compute the purpose distribution as the fraction of citation events assigned to each purpose label:

$$P_l = \frac{N_l}{N},$$

where N_l is the number of events with purpose label l , and N is the total number of citation events in the paper.

Stance Distribution. Each citation event is also assigned one of three stance labels. *Positive* indicates that the cited work is used positively or as support for the citing claim. *Negative* indicates that the citing text challenges, criticizes, or disagrees with the cited work. *Neutral* indicates a neutral citation without a clear supportive or contrastive attitude. We compute stance distribution analogously as the fraction of citation events assigned to each stance label:

$$S_l = \frac{N_l}{N}.$$

B.3 Citation Topology Metrics

This section provides implementation details for the citation topology metrics introduced in Section 3.3. These metrics characterize structural properties of a reference list, including repetitiveness, prominence, recency, and diversity. Each metric is computed per paper and then averaged across the 50 papers per condition.

Let p denote a target paper (either a survey or a method paper), and let $\mathcal{R}$ denote the set of reference list entries of p matched to OpenAlex, either by DOI lookup or by title-based search where a match is accepted only when the normalized character overlap between the reference title and the retrieved title exceeds 0.9. Let $|\mathcal{R}|$ be the number of matched references. The OpenAlex reference matching rate averaged $83.79\% \pm 4.26\%$ across conditions, indicating consistently high coverage.

In-Text Repetition Rate. This captures the tendency to repeatedly cite the same work within a paper. The in-text repetition rate is defined as:

$$\text{In-Text Repetition Rate} = 1 - \frac{|\mathcal{R}|}{C},$$

where c_r denotes the in-text citation count of reference $r \in \mathcal{R}$, and $C = \sum_{r \in \mathcal{R}} c_r$ is the total number of in-text citation events. A higher value indicates that a small number of references are cited repeatedly (Algaba et al., 2025b).

Citation Count. This metric reflects the prominence of the cited works. For each $r \in \mathcal{R}$, we collect its cumulative citation count κ_r from OpenAlex (Priem et al., 2022). The per-survey score is the median citation count across $\mathcal{R}$:

$$\text{Citation Count} = \text{median}_{r \in \mathcal{R}} \{\kappa_r\}.$$

A higher value indicates that the reference list is dominated by highly cited, well-known works (Algaba et al., 2025b).

Years Gap. This metric captures the recency of cited works relative to the target paper. Let y_p denote the publication year of p and y_r the publication year of reference $r \in \mathcal{R}$. The years gap follows:

$$\text{Years Gap} = \text{median}_{r \in \mathcal{R}} \{y_p - y_r\}.$$

A higher value indicates a tendency to cite older works (Bao et al., 2025).

Venue Diversity. This metric measures the variety of publication venues across the reference list. Let $v(r)$ denote the publication venue of reference r in OpenAlex. The venue diversity ratio is:

$$\text{Venue Diversity} = \frac{|\{v(r) : r \in \mathcal{R}\}|}{|\mathcal{R}|}.$$

A higher value indicates that the cited works are drawn from a wider variety of publication venues (Algaba et al., 2025b).

Field Diversity. This metric measures the variety of research areas across the reference list. Let $s(r)$ denote the OpenAlex subfield assigned to reference r . The field diversity ratio is:

$$\text{Field Diversity} = \frac{|\{s(r) : r \in \mathcal{R}\}|}{|\mathcal{R}|}.$$

A higher value indicates that the cited works span a broader range of research areas (Yegros-Yegros et al., 2015).

Clustering Coefficient. This metric captures how cohesively the cited works are interconnected in the citation graph. Let $G_p = (\mathcal{R}, E)$ denote the graph induced by the references cited in paper p , where an edge connects two references if one cites the other. For each reference r , we compute its local clustering coefficient and average it across $\mathcal{R}$:

$$\text{Clustering Coefficient} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{2e_r}{k_r(k_r - 1)},$$

where k_r is the degree of r and e_r is the number of edges among its neighbors. A higher value indicates a more interconnected reference set. Unlike the other topology features, which are computed from the directly cited references, the clustering coefficient additionally uses 2-hop citation information to capture citation links among references.

C Dataset Details

C.1 Dataset Statistics

CITESPECTRUM contains 550 papers and 78,016 citation events. The survey subset comprises 450 papers and 73,586 citation events, including 400 AI-generated surveys and 50 human-authored surveys. The AI-generated surveys are produced using two models (GPT-5.4 and Qwen3-80B) under four generation settings (LLM-Only, RAG, Outline-Guided, and Text-Guided), with 50 papers for each model–setting combination. The same 50 human-authored surveys serve as the reference condition for both models. The method subset comprises 100 papers and 4,430 citation events with 50 AI-generated and 50 human-authored papers.

The survey subset spans 14 topical domains: Psychology, Computer Science, Biology, Medicine, Chemistry, Engineering, Geography, Materials Science, Physics, Economics, Sociology, Mathematics, History, and Business. Each source topic corresponds to nine papers: eight AI-generated surveys (two models $\times$ four settings) and one human-authored survey. The first four domains (Psychology, Computer Science, Biology, Medicine) contain ten source papers each, while each of the remaining ten domains contains one, resulting in 90 and 9 papers per domain, respectively.

Figure 5 presents the paper-level distribution of length-normalized citation density. For each paper, citation density is calculated as the number of citation events per 1,000 body-text words, excluding the reference list. Among GPT-5.4 surveys, the median densities are 24.28 [16.63–30.44] for LLM-Only, 26.37 [20.70–31.56] for RAG, 32.89 [25.86–37.05] for Outline-Guided, and 27.19 [16.87–33.81] for Text-Guided, where brackets denote the interquartile range. The corresponding Qwen3-80B densities are 20.35 [14.60–25.89], 11.84 [9.59–14.43], 19.51 [16.76–21.99], and 22.32 [13.45–29.98], respectively. Human-authored surveys have a median density of 23.62 [17.99–32.20] citation events per 1,000 words. Additionally, method papers have lower citation density than survey papers

due to their less literature-oriented structure.

Overall, AI-generated and human-authored surveys show broadly similar citation densities, indicating that citation frequency is comparable across authorship conditions after controlling for document length. Within the AI-generated surveys, GPT-5.4 generally exhibits higher citation density than Qwen3-80B, showing a consistent generator-level difference in citation behavior across several settings. At the same time, variation remains within each condition, which is expected given the diversity of papers, domains, and publication venues represented in the dataset.

C.2 Document Processing and Citation Extraction

We first convert each paper PDF into structured text using PaddleOCR-VL (Cui et al., 2025), a state-of-the-art document parsing model reporting approximately 97.9% text-recognition accuracy on academic literature. The OCR output is used to extract paper content, reference lists, citation markers, and citation sentences. We then identify citation markers using rule-based patterns that cover common numeric and author-year citation formats. For example, we detect numeric citations such as “[1]”, “[3, 5]”, and “[7–9]”, as well as author-year citations such as “Smith et al. (2020)”, “(Smith and Lee, 2021)”, and “Smith et al., 2020”. These rules provide an initial set of citation markers and their corresponding citation sentences.

For citation instances that are not reliably captured by the rule-based extractor, we use GPT-5.4-nano to recover citation markers from the OCR-parsed text. The model is given the local sentence context and asked to return only explicit citation markers appearing in the text. This second stage is used only as a fallback to improve recall for noisy OCR outputs and irregular citation formats.

We parse the reference list from the OCR text and extract structured metadata, including reference number, title, authors, venue, and year when available. For numeric citations, we match citation markers to references using the reference number. For author-year citations, we match citation markers to references by normalizing author surnames and publication years, and then linking them to the corresponding parsed reference entry. Citation instances that cannot be matched to a parsed reference are excluded from claim–reference pair construction.

Finally, for each matched reference, we re-

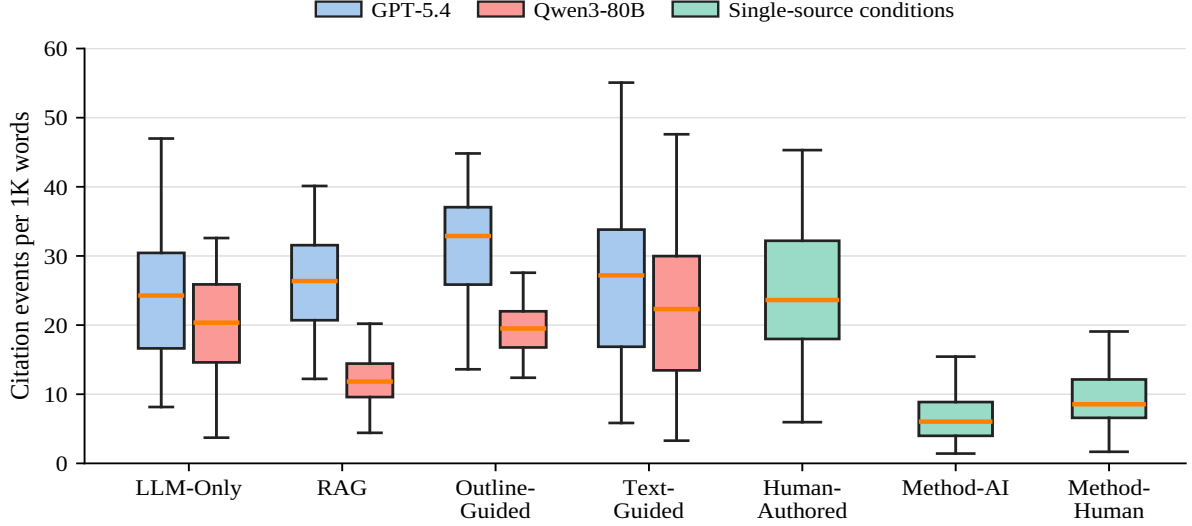


Figure 5: Boxplot distribution of citation density (citation events per 1,000 words) in CITESPECTRUM. Each condition contains 50 papers. Boxes show Q1–Q3 with median lines (orange) and 1.5×IQR whiskers.

trieve its full text when available through Google Scholar, Semantic Scholar, OpenAlex, arXiv, publisher pages, and web search using GPT-5.4-nano. Overall, full text was obtained for 80.2% of successfully matched references, with coverage ranging from 75.6% to 82.8% across conditions; references without available full text were excluded from evidence-grounding evaluation. The retrieved PDFs are processed with the same OCR pipeline, and the resulting text is segmented into paragraphs for evidence retrieval, evidence-grounding labeling, and provenance analysis.

D Experimental Details

D.1 Detection Data

The AI/Human detection experiment uses the corpus introduced in Section 3.1, but excludes the LLM-Only survey condition because its high ghost-reference rate makes citation-based feature extraction unreliable beyond reference-validity analysis. The resulting classification set contains 450 documents: 300 AI-generated surveys from RAG, Outline-Guided, and Text-Guided; 50 human-authored surveys; 50 AI-generated method papers; and 50 human-authored method papers.

As an additional screening step, we evaluate the 100 human-authored documents using Pangram; all are classified as human-authored under our screening criterion.

D.2 Citation-MLP Configuration

Our feature-based classifier is a multi-layer perceptron trained on 16 paper-level citation-behavior features instantiated from the 11 feature groups defined in Section 3.3. The 16-dimensional feature vector includes three fidelity features, seven usage features, and six topology features. Fidelity features include supported citation ratio, introduction provenance score, and body provenance score. Usage features include four purpose ratios and three stance ratios. Topology features include median reference age gap, venue diversity, field diversity, median citation count, in-text repetition rate, and clustering coefficient. We remove the unsupported ratio and the result provenance score, which are linearly determined by the remaining fidelity features, and the bundling rate, which we found unstable across the survey and method domains.

The MLP uses two hidden layers with sizes 52 and 36, ReLU activation, and standard feature normalization. We train using the Adam optimizer with $\text{max_iter}=3000$ and convergence tolerance 10^{-4} . The architecture is selected from a grid of one-layer and two-layer MLPs whose hidden sizes are multiples of four, using Balanced Accuracy as the selection criterion.

D.3 Evaluation Details

Fold Construction. We use source-grouped, field-balanced 5-fold cross-validation. The 50 source surveys are divided into five balancing groups—Psychology, Computer Science, Biology,

Medicine, and Others—each containing 10 source surveys. The first four groups correspond to single domains, whereas the Others group contains one survey from each of the remaining ten domains. Each fold holds out two source surveys from each group, yielding 10 held-out source surveys per fold; within Others, the two held-out surveys come from distinct domains, distributing its ten constituent domains across the five folds. All documents derived from the same source survey are assigned to the same fold. In the detection setting, each source survey contributes six AI-generated papers (three generation settings across two generators) and one Human-Authored paper, so each survey test fold contains 60 AI-generated and 10 Human-Authored papers. Method papers are assigned to folds by the same source identifiers, contributing 10 AI-generated and 10 Human-Authored papers to each fold. Each test fold therefore comprises 90 documents. This construction prevents source-topic leakage between the training and test sets.

Evaluation Setup. We fix the classification threshold at 0.5 for all threshold-based metrics and report Balanced Accuracy, Macro-F1, and AU-ROC. Zero-shot baselines are evaluated on the full 450-sample dataset without cross-validation. Cross-validation is applied only to supervised models (Citation-MLP, DistilBERT, and Llama-3.2-3B).

D.4 Baseline Configurations

We compare CITATION-MLP against off-the-shelf and in-domain supervised text-based baselines. All experiments are conducted on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (96 GB VRAM each).

Qwen3-Next. We use Qwen3-Next-80B-A3B-Instruct as a zero-shot classifier. The model is prompted to classify each paper as AI-written or human-written, and we extract next-token logits over the 0 and 1 labels, thresholding at 0.5. Inference is performed in float16 precision with thinking mode disabled. Inference requires 2 GPUs and completes in approximately 30 minutes.

Fast-DetectGPT. Fast-DetectGPT is a zero-shot detector that scores each text by the discrepancy between its conditional log-probability under a scoring model and samples drawn from a reference model. We use falcon-7b as the reference model and falcon-7b-instruct as the scoring model, with maximum token length of 2048, threshold-

ing scores at 0.5. Inference requires 1 GPU and completes in approximately 15 minutes.

EditLens. EditLens is a zero-shot AI-text detector based on Llama-3.2-3B with a LoRA adapter fine-tuned by Pangram. We run inference with a maximum context length of 131,072 tokens (full text) in 4-bit quantization. Inference requires 1 GPU and completes in approximately 5 minutes.

Pangram. Pangram is a commercial AI-text detector. Since Pangram returns fractional predictions over categories such as AI, AI-assisted, Human, and Mixed, we map documents with fraction-human greater than 0.5 to the human class and all others to the AI class.

Llama-3.2-3B. We fine-tune Llama-3.2-3B for binary AI/Human classification using LoRA adapters in bfloat16 precision. Rather than generating label tokens, we attach a sequence classification head that projects the last-token hidden state to two logits, implemented as a LayerNorm followed by a bias-free linear layer. LoRA adapters (rank 8, scaling factor 16, dropout 0.05) are applied to all attention and feed-forward projections, while the classification head is trained in full. The model is trained with learning rate 3×10^{-5} for 4 epochs. To handle full documents, each text is split into consecutive 4096-token chunks that share the document label. At inference, the per-chunk human probabilities are averaged into a single document-level score. We apply the same citation-marker removal as DistilBERT for survey texts, and remove author blocks and emojis for method papers. All texts are normalized with NFKC normalization and lowercased. Training and evaluation across five folds require 1 GPU and complete in approximately 5 hours.

DistilBERT. We fine-tune distilbert-base-uncased for binary AI/Human classification using DistilBertForSequenceClassification with a single logit output. We optimize BCEWithLogitsLoss with label smoothing of 0.1. The model is trained with learning rate 2×10^{-5} , batch size 16, nine epochs, maximum token length 256, AdamW optimization, and gradient clipping at 1.0. To reduce superficial citation-format cues, we remove citation markers from survey texts, including ref[n], bracketed numeric citations, and parenthetical content. For method papers, we remove author blocks and emojis. We normalize all texts with NFKC and lowercase them. To account for class imbalance, we use pos_weight=3.5. Training and

Setting	GPT-5.4	Qwen3-80B
Backend	OpenAI API	Local vLLM
Model	gpt-5.4	Qwen3-Next-80B
Reasoning / Thinking	none	Disabled
Temperature	1.0	1.0
Context Length	1.05M	32,768
Max Generation	128k [†]	8,192
Tensor Parallelism	–	2
GPU Memory Util.	–	0.90

Table 7: Generator configurations for AI-involved survey generation. Lengths are reported in tokens. [†]No explicit output limit was set for GPT-5.4; 128k denotes the model’s maximum supported output length.

evaluation across five folds require 1 GPU and take about 5 minutes.

E Generation Details

This section summarizes the generator configurations and prompt templates used to construct the AI-involved survey conditions. Curly-braced fields such as {topic} denote dynamic values filled at runtime. Across all conditions, the model is instructed to return structured JSON output, allowing us to parse generated paragraphs and reference lists consistently. When candidate references are provided, references are indexed and the model is instructed to cite them using the format ref [n]. We present the prompts in increasing order of human guidance: LLM-Only, RAG, Outline-Guided, and Text-Guided. Condition-specific citation constraints are included directly inside the corresponding prompt templates.

E.1 Generator Configuration

We used two generators for the AI-involved survey conditions: GPT-5.4 through the OpenAI backend and Qwen3-Next-80B-A3B-Instruct through a local vLLM backend. Table 7 summarizes the user-visible generation settings. For GPT-5.4, we used reasoning_effort=none. We did not set a separate maximum output length. For Qwen3-Next-80B-A3B-Instruct, we used a maximum context length of 32,768 tokens and a maximum generation length of 8,192 tokens. We used temperature=1.0 for survey generation with both generators. We did not observe truncation or generation failures attributable to these token limits in the final parsed survey outputs.

E.2 LLM-Only Generation

The LLM-Only condition does not provide an external reference list. The model generates both the sur-

vey outline and subsection content from its parametric knowledge, while being explicitly instructed not to fabricate citations. The outline and subsection writing prompts are shown in Tables 8 and 9. The subsection writing prompt includes the LLM-Only-specific constraint that the model should cite only real, existing scholarly papers and should omit citations when it is not confident that a paper exists.

E.3 RAG Generation

The RAG condition uses externally retrieved candidate references rather than the original human reference list. The retrieved pool is constructed using the Semantic Scholar API with topic- and title-based queries, followed by expansion through related-paper recommendations. We remove duplicates and match the final pool size to the reference budget of the corresponding human survey. We use three steps: outline generation, reference preselection, and subsection writing. During subsection generation, the model first preselects references using titles and abstracts; the full texts of only the selected references are then provided to the model for writing. The corresponding prompts are shown in Tables 10, 11, and 12. The subsection writing prompt includes the RAG-specific constraint that citations must be selected only from the provided candidate reference list.

E.4 Outline-Guided Generation

The Outline-Guided condition uses the human survey’s section structure and human-selected reference pool, but the model generates the prose. We first organize each section into subsections and map candidate references to the corresponding subsections. The subsection planning prompt is shown in Table 13. Subsection writing uses the reference-conditioned template in Table 12, with the Outline-Guided-specific writing policy in Table 14. This policy modifies the citation rule to prioritize the human-selected reference pool and instructs the model to use a diverse subset of mapped section references.

E.5 Text-Guided Generation

The Text-Guided condition starts from human-written survey prose with citation markers removed. The model receives the citation-erased subsection text and a mapped candidate reference list, and is asked to restore suitable in-text citations while preserving the original prose. The citation restoration prompt is shown in Table 15.



Figure 6: Boxplot distributions of citation fidelity subfeatures across AI involvement levels. Each plot compares feature values across RAG, Outline-Guided, Text-Guided, and Human-Authored survey papers, with an additional binary AI/Human comparison for method papers. Red dashed lines connect group means.

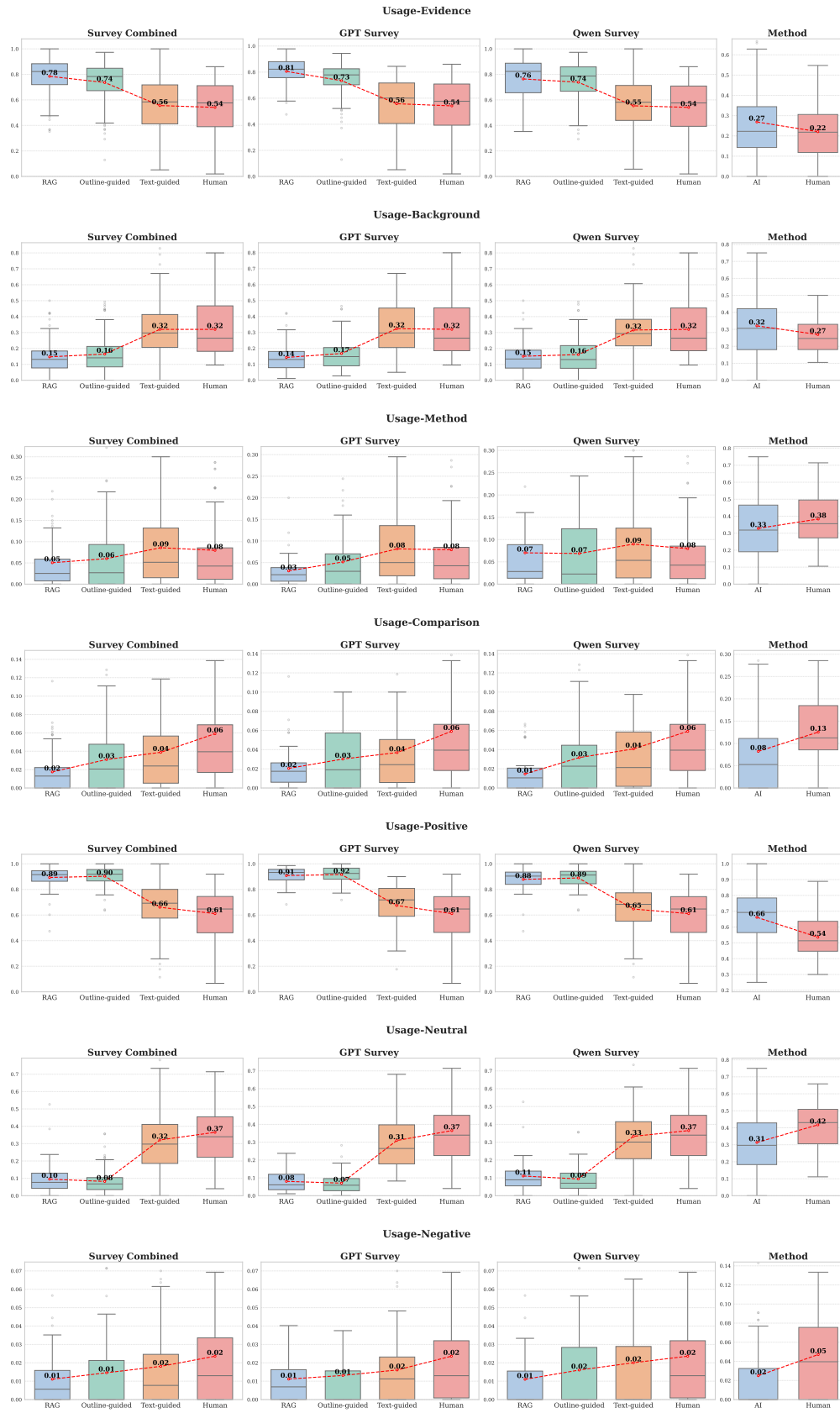


Figure 7: Boxplot distributions of citation usage subfeatures across AI involvement levels. Each plot compares feature values across RAG, Outline-Guided, Text-Guided, and Human-Authoring survey papers, with an additional binary AI/Human comparison for method papers. Red dashed lines connect group means.

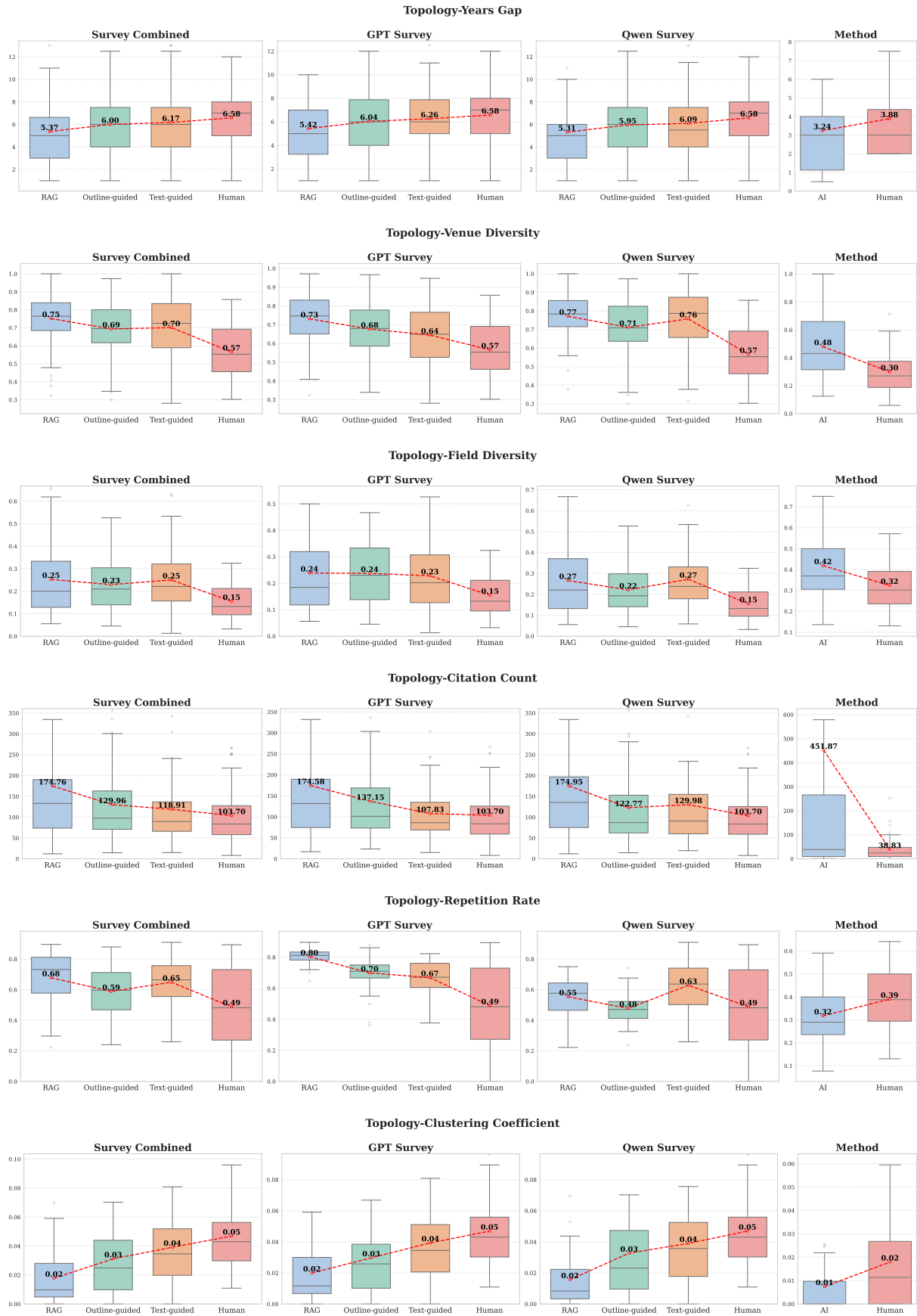


Figure 8: Boxplot distributions of citation topology subfeatures across AI involvement levels. Each plot compares feature values across RAG, Outline-Guided, Text-Guided, and Human-Authored survey papers, with an additional binary AI/Human comparison for method papers. Red dashed lines connect group means.

LLM-Only Outline Generation Prompt

User prompt:

You are an academic expert familiar with the following research topics:

{topic}

Your task is to generate a well-structured outline for a survey that covers these topics. You have no paper titles or abstracts available – only the list of topics above. Use them creatively to identify major themes and subtopics appropriate for a survey.

Please follow the instructions below:

Step 1: Based solely on the provided topics, identify 3 to 7 major thematic sections that define the overall scope and objectives of the survey. For each section, provide an academically styled title, along with a brief description summarizing its focus and relevance.

Step 2: For the major thematic sections, list several subsections representing more specific research areas, concepts, or points to be covered.

Step 3: Your output should be in JSON format and must include the survey title, a structured outline with section titles, descriptions, and subsections.

Return only the structured JSON output.

Table 8: Prompt for generating an outline in the LLM-Only condition.

LLM-Only Subsection Writing Prompt

User prompt:

You are an academic expert familiar with the research themes indicated by the survey outline below. Your task is to write a subsection of a survey for the following topic area using any relevant knowledge you possess.

The overall structure of the survey is as follows:

- {outline_text}

You are now asked to write the following subsection:

- Subsection: {subsection_title}

No specific reference list is provided in this mode. Therefore, follow the **LLM-Only citation constraint**:

- Cite ONLY real, existing scholarly papers (published or widely recognized preprints).

- Do NOT invent, hallucinate, or guess references.

- If you are not confident a paper exists, do not cite it.

Instructions for generating subsection content:

1. Ensure the content directly addresses the specific topic of the subsection. Aim for a concise write-up (~150–200 words).

2. Maintain alignment with the parent section and overall survey topic.

3. When a statement specifically depends on prior literature, use in-text citations in the format ref [1], ref [2], etc.

4. Every citation must correspond to a real paper title. Never output placeholder, hypothetical, or fabricated references.

5. Use a formal academic tone.

6. Return STRICT JSON only:

```
{
  "paragraphs": ["paragraph 1 ...", "paragraph 2 ..."],
  "references": [{"refNo": 1, "title": "title of the reference"}]
}
```

Return ONLY the JSON object. No markdown fences, no commentary.

Table 9: Prompt for generating subsection content in the LLM-Only condition.

RAG Outline Generation Prompt

User prompt:

The following highly relevant papers, including their titles and abstracts, are provided for reference:

`{references_list}`

Please follow the instructions below:

Step 1: Based on the given topic, identify 3 to 7 major thematic sections that define the overall scope and objectives of the survey. For each section, provide an academically styled title, along with a brief description summarizing its focus and relevance.

Step 2: For the major thematic sections, list several subsections representing more specific research areas, concepts, or points to be covered. Subsections should be conceptually related to their parent section and serve to further structure the survey.

Step 3: Your output should be in JSON format and must include the survey title, a structured outline with section titles, descriptions, and subsections. Ensure the structure is logically coherent, well-aligned with the topic, and suitable for developing a full-length academic survey.

Return only structured JSON.

Now, based on the given topic `{topic}`, please generate the outline.

Table 10: Prompt for generating an outline in the RAG condition.

RAG Reference Preselection Prompt

User prompt:

You are selecting references for subsection writing.

Topic: `{topic}`

Section context:

`{section_summary}`

Target subsection: `{subsection_title}`

Candidate references (title only):

`{refs_block}`

Task:

1. Select only references that are likely essential for this subsection.
2. Use both the reference title and abstract/snippet to judge relevance.
3. Prioritize references that directly match the survey topic, section context, and subsection topic.
4. Avoid tangential references that are only loosely related.
5. Keep selection concise to reduce context size.
6. Return up to `{max_pick}` reference numbers.

Return STRICT JSON only:

```
{
  "selected_ref_nos": [1, 2, 5]
}
```

Table 11: Prompt for selecting subsection-level references in the RAG condition.

Reference-conditioned Subsection Writing Prompt for RAG
User prompt: You are an academic expert in the field of “{topic}” with deep expertise in survey writing. Your task is to write a subsection of a survey, based on the following information: Section info: {section_summary} You are now asked to write the following subsection: - Subsection: {subsection_title} The following reference candidates are provided (title-based list): {refs_prompt} {citation_guidance} Instructions for generating subsection content: 1. Carefully analyze the provided references and identify those highly relevant to the subsection topic as the basis for your generation. 2. Ensure the content directly addresses the specific topic of the subsection. Aim for a concise write-up (~150–200 words); shorter subsections are preferred to reduce token usage. 3. Ensure key literature-dependent claims are supported by relevant academic references. Use in-text citations in the format ref [1], ref [2], etc. Generic bridge/transition sentences do not need citations. 4. Citation policy: - Use ONLY references from the provided candidate list. - Do not introduce uncited or out-of-list papers. - Never fabricate references. Every cited paper must be real. 5. When candidate entries include long full text, focus on the main body; if a trailing “References/Bibliography” list appears inside that text, ignore it for evidence/content understanding. 6. Maintain alignment with the parent section and overall survey topic. 7. Use a formal academic tone. 8. Return STRICT JSON only: <pre>{ "paragraphs": ["paragraph 1 ...", "paragraph 2 ..."], "references": [{ "refNo": 1, "title": "title of the reference" }] }</pre> Return ONLY the JSON object. No markdown fences, no commentary.

Table 12: Reference-conditioned subsection writing prompt used in the RAG condition.

Outline-Guided Subsection Planning Prompt
User prompt: You are organizing an academic survey section into concise subsections. Survey title: {survey_title} Section title: {section_title} Section description: {section_description} Candidate references for this section: {reference_titles_and_abstract_snippets} Task: 1) Create between {min_subsections} and {max_subsections} subsection titles for this section (not too many). 2) The subsection titles must stay tightly aligned with the survey title and section title. 3) Assign candidate references to each subsection by index. 4) Prefer references whose titles/abstracts are directly relevant to the survey topic and subsection topic. 5) Avoid generic or weakly related references even if they mention the broad topic. 6) Do not create broad catch-all subsections like “Background” unless the section title itself is already specific. 7) Use each reference at most once if possible, to reduce prompt redundancy. 8) Only use the provided reference indices; do not invent references. Return ONLY valid JSON in this format: <pre>{ "subsections": [{ "title": "subsection title", "description": "brief focus", "ref_indices": [1, 2, 3] }] }</pre>

Table 13: Prompt for constructing subsection-level plans in the Outline-Guided condition.

Outline-Guided Subsection Writing Prompt
The Outline-Guided condition uses the same reference-conditioned subsection writing template as Table 12, but with the following softer citation policy because the candidate pool is derived from human-selected references for the corresponding survey section. 4. Citation policy: - PRIORITIZE and prefer papers from the provided reference list. - You MAY additionally cite other real, highly relevant papers when needed for coverage/completeness. - Never fabricate references. Every cited paper must be real. 5. Diversity constraint (human-guided): - For each subsection, select a varied subset from the mapped section references instead of repeatedly citing the same few titles.

Table 14: Modified citation policy used for subsection writing in the Outline-Guided condition.

Text-Guided Citation Restoration Prompt

User prompt:

You are given citation-erased survey prose and a mapped reference list.

Your task is to restore suitable in-text citations (ref [n]) by matching claims to the provided references. Preserve the original meaning and structure.

Topic: {topic}

Section/Subsection context:

{outline_text}

Target subsection:

- {subsection_title}

Citation-erased source text to revise:

{source_text}

Reference list (use only these references for in-text citations):

{refs_prompt}

{citation_guidance}

Rules:

1. Keep the source text wording as-is as much as possible. Do NOT paraphrase or rewrite unless minimal edits are required to place citations.
2. Insert citations in the format ref [1], ref [2], ... at appropriate claim-level positions.
3. Use ONLY references from the provided list; do not introduce external papers.
4. If a sentence is generic and not literature-dependent, citation may be omitted.
5. Preserve original sentence order and paragraph structure from the provided source text.
6. Return STRICT JSON only:

```
{
  "paragraphs": ["paragraph 1 ...", "paragraph 2 ..."],
  "references": [{"refNo": 1, "title": "..."}]
}
```

Return ONLY the JSON object.

Table 15: Prompt for restoring citations in the Text-Guided condition.

Evidence Grounding Labeling Prompt

User prompt:

You are an academic citation checker. Decide whether the reference paragraph supports the given claim.

Claim:

{claim}

Reference Paragraph:

{reference_paragraph}

Choose one label:

- supported: The paragraph provides direct, partial, or reasonable evidence for the claim.
- unsupported: The paragraph does not support the claim, is only topically related, or contradicts the claim.

Judge only based on the given paragraph. Do not use outside knowledge.

Return STRICT JSON only:

```
{
  "label": "supported" or "unsupported",
  "reason": "brief explanation"
}
```

Return ONLY the JSON object.

Table 16: Prompt for labeling whether a retrieved reference paragraph supports the corresponding claim.

Citation Usage Labeling Prompt
System prompt: You are labeling citation behavior in scientific writing. Return strict JSON only. User prompt: Task: label BOTH citation purpose and stance for this citation event. Purpose labels: - background: provides background/context/general prior knowledge - method: directly uses/adopts method, model, data, software, concept (e.g., “we use/follow”) - comparison: contrasts or compares approaches/results (if table/figure/caption context, comparison can be preferred) - evidence: supports/justifies an explicit claim or interpretation Stance labels: - positive: citing author is positive/supportive toward cited work/claim - negative: citing author challenges/criticizes/disagrees - neutral: neutral mention without clear support/contrast Event: - section_title: {section_title} - section_bucket: {section_bucket} - is_bundle: {is_bundle} - target_marker_group: {marker_group} - citing_sentence: {citing_sentence} - citation_block(prev+current+next): {block_text} - linked_references: {linked_references_formatted} Disambiguation rule: - If multiple citation markers appear in one sentence, label ONLY the citation event for target_marker_group and its linked references, rather than other markers in the same sentence. Output JSON schema: <pre>{ "purpose_label": "one of ['background', 'method', 'comparison', 'evidence']", "stance_label": "one of ['positive', 'negative', 'neutral']", "confidence": 0-100 }</pre>

Table 17: Prompt for labeling citation purpose and stance.