ReCite: Agentic Reasoning for Faithful Citation

Yuyang Huang[1], Bobo Li[2]*, Jiajia Song[2],
Yuzhe Ding[1], Chong Teng[1], Fei Li[1], Donghong Ji[1]*
[1]School of Cyber Science and Engineering, Wuhan University
[2]National University of Singapore
{hyy279, dhji}@whu.edu.cn, {libobo, jjiasong}@nus.edu.sg

Abstract

Accurate citations are the foundation of academic writing, tracing intellectual origins and substantiating core claims. However, manually navigating the growing volume of scientific literature is increasingly difficult, prompting reliance on automatic citation recommendation. While modern retrieval-augmented architectures have largely mitigated the fabrication of non-existent papers, current systems relying on semantic similarity struggle with misattribution, often citing authentic papers that fail to logically support the author's claim. To address this challenge, we argue that accurate citation requires a shift from similarity-based search to active, claim-level reasoning. We propose ReCite, a decoupled agentic framework that orchestrates location perception, intent-aware query planning, and reflective verification. Trained on synthesized reasoning trajectories, our agent verifies claim-evidence consistency and triggers self-correction loops when retrieved candidates lack logical support. Experiments demonstrate that our lightweight framework outperforms state-of-the-art massive generative models in strict citation accuracy. By grounding literature matching in verifiable logic rather than semantic overlap, ReCite establishes a reliable foundation for automated academic writing. Our code and data are available at https://github.com/Hyy279/ReCite.

1 Introduction

Accurate citations are the foundation of academic writing, serving to trace intellectual origins and substantiate core claims. However, as the volume of scientific literature grows exponentially, manually tracking relevant prior work and locating the most appropriate papers have become increasingly difficult (Bornmann and Mutz, 2015; Fortunato et al., 2018). To alleviate this growing cognitive burden, automatic citation recommendation (ACR) has been widely adopted as an essential research tool (Gao et al., 2023). As large language models (LLMs) increasingly draft academic texts, they introduce a severe risk of generating plausible but incorrect references (Ji et al., 2023; Sakai et al., 2026), forcing ACR to evolve from a passive writing aid into an active quality gatekeeper.

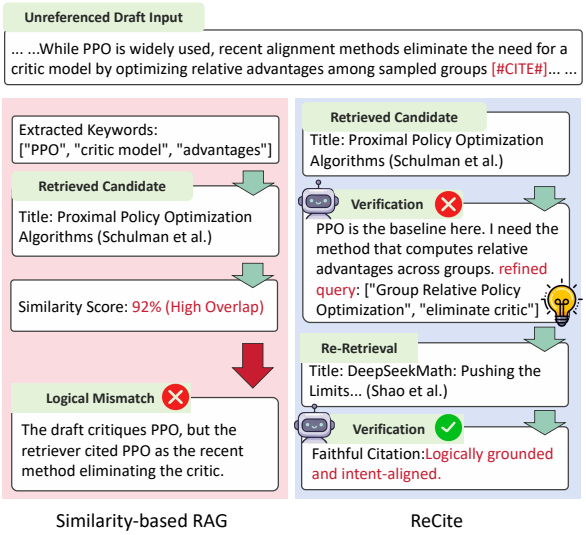

Figure 1: Comparison of citation paradigms: Traditional semantic retrieval versus our framework.

For years, citation recommendation has been modeled as a similarity search problem (Gu et al., 2022; Gao et al., 2023; Wang et al., 2025). Both early retrievers (Huang et al., 2014) and recent LLM-based agents (Çelik and Tekir, 2025; Zhang et al., 2026) equate citing with finding a textually relevant paper. In the LLM era, citation hallucinations (Algaba et al., 2025; Sakai et al., 2026) have crystallized into two distinct forms. Fabrication, inventing non-existent papers, is largely resolved by retrieval-augmented architectures; misattribution, citing authentic papers that fail to logically support the claim, exposes the core limit of similarity-based retrieval and remains open (Xu et al., 2025).

*Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 31441–31454
October 24-29, 2026 ©2026 Association for Computational Linguistics

To address this, we argue that citation is fundamentally a reasoning task, not a retrieval one: topical relevance does not guarantee logical support for a claim. For example, a claim about the Transformer's success in NLP cannot be supported by the vision transformer paper (Dosovitskiy et al., 2021) despite their textual overlap; "Attention is All You Need" (Vaswani et al., 2017) is the correct logical pillar. Yet semantic retrievers (Cohan et al., 2020) ignore that different citation contexts, such as introducing a background versus comparing a method, demand distinct verification criteria (Cohan et al., 2019). Worse, current systems operate as unidirectional pipelines (Medić and Snajder, 2020), so a single faulty retrieval becomes an irrevocable error without opportunity for self-correction. Together, solving misattribution requires intent-aware retrieval, claim-level verification, and a reflective loop for error recovery.

To fulfill these requirements, we propose ReCite (Figure 1), an agentic reasoning framework for faithful citation, trained via supervised fine-tuning and reinforcement learning. It reshapes the citation process from a linear pipeline into an iterative architecture of reasoning, retrieval, verification, and reflection. Operating in an observe-think-act format (Yao et al., 2023), the agent is built on Qwen3-4B (Yang et al., 2025), with key modules reinforced via Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Within this framework, we instantiate three core mechanisms. First, claim-evidence verification tackles misattribution by cross-checking each candidate's metadata against the inferred citation intent, accepting a citation only upon consistent evidence. Second, an intent-aware query planner addresses context misalignment by generating retrieval keywords conditioned on predicted citation intents. Third, a reflective re-retrieval loop overcomes single points of failure. Upon verification failure, the agent analyzes the cause, rewrites the query, and retries. Since retrieval queries only target authentic databases, hallucinating entirely non-existent papers is eliminated; ReCite then focuses entirely on resolving misattribution.

Experiments on the benchmark show that ReCite substantially outperforms strong generative baselines in strict citation accuracy. Ablations confirm that claim-evidence verification, intent-aware planning, and the reflective loop are indispensable to this gain. In summary, our main contributions are:

- We systematically decouple citation hallucinations into fabrication and misattribution, establishing misattribution as the core challenge in the LLM era and reframing citation as a reasoning task.
- We propose ReCite, the first closed-loop claim-evidence verification framework, integrating intent-aware query planning and a reflective re-retrieval loop.
- Experiment results highlight the inherent challenge of citation reasoning for current LLMs and demonstrate the effectiveness of ReCite.

2 Related Work

Local and Global Citation Recommendation. ACR has transitioned from global representations and two-stage pipelines (Beltagy et al., 2019; Medić and Snajder, 2020; Gu et al., 2022) to localized, context-aware generation (Çelik and Tekir, 2025; Buscaldi et al., 2024) and dynamic Retrieval-Augmented Generation (RAG) frameworks (Wang et al., 2025; Zhang et al., 2026). Although these generative models excel at contextual integration, relying on semantic similarity as a generative prior often bypasses strict logical entailment. Consequently, these systems frequently suffer from citation fidelity issues, including hallucinated references and misattributed claims (Algaba et al., 2025; Sakai et al., 2026; Çelik and Tekir, 2025).

Logical Verification and Agentic Reasoning. Standardized benchmarks such as ALCE (Gao et al., 2023) have begun to quantify citation reliability, prompting verification approaches based on Natural Language Inference (NLI) (Xu et al., 2025) or theorem provers (Pei et al., 2025). However, such static post-hoc filtering cannot correct errors that have propagated through a unidirectional generation pipeline. Autonomous frameworks (e.g., ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023)) and step-level reinforcement (e.g., GRPO (Li et al., 2026b)) introduce self-correction (Li et al., 2026a), but remain constrained by pre-trained priors and incur high context overhead. Current citation datasets only map local contexts to target papers; they omit the explicit reasoning traces required to train agents for this task. To build reasoning-driven citation models, the field requires a specialized dataset that supports location perception, intent reasoning, and closed-loop reflective trajectories.

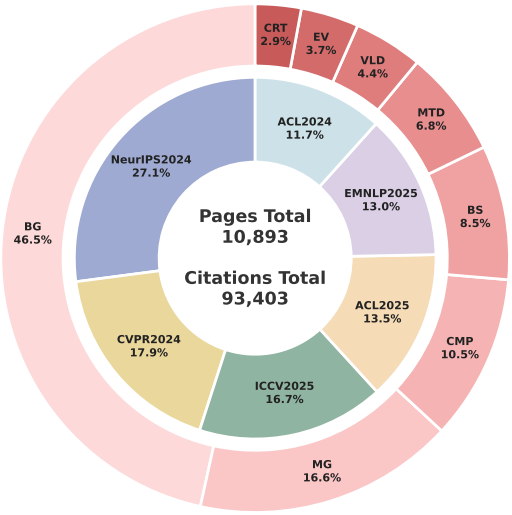

Figure 2: The proportion distribution of academic conference sources and the various citation intent categories within the utilized citation dataset.

3 Dataset Construction

We construct a large-scale literature dataset for training an agentic citation workflow. Given a draft paragraph, the agent must (1) detect which positions require a citation, (2) infer each citation's intent and generate targeted retrieval queries, and (3) verify whether retrieved candidates support the local claim, retrying with refined queries on failure. The dataset therefore provides three task-specific subsets that supply the supervision for the three corresponding modules of ReCite, namely CiteLocator, QueryPlanner, and Master Brain.

3.1 Data Source and Preparation

Our raw corpus consists of 10,893 LaTeX source packages crawled from arXiv, corresponding to papers published in 2024–2025 at top computer science venues, including ACL, EMNLP, CVPR, ICCV, and NeurIPS. Working at the LaTeX source level preserves the explicit `\cite` commands and their bibliography keys, giving us ground-truth citation positions and references for downstream supervision.

We apply a four-step preparation pipeline: extracting introduction sections, normalizing heterogeneous citation macros (`\cite`, `\citep`, `\citet`, etc.) into a unified form, filtering non-textual noise, and removing paragraphs with abnormal citation density. Figure 2 shows the venue-level distribution of the resulting corpus. Compared with existing citation datasets such as RefSeer (Huang et al., 2014) and S2ORC (Lo et al., 2020), which provide only static context-to-reference mappings, our dataset additionally captures the location, intent, and reflective trajectories required to train an agentic citation workflow (Table 1).

Features	RefSeer	S2ORC	Ours
Local Context Mapping	✓	✓	✓
Full-Text Structure	✗	✓	✓
Location Perception	✗	✗	✓
Intent Reasoning	✗	✗	✓
Reflective Trajectories	✗	✗	✓

Table 1: Comparison of our constructed trajectory dataset with existing citation datasets, including RefSeer (Huang et al., 2014) and S2ORC (Lo et al., 2020).

Category	Abbr.	Description
Background	BG	Provides background knowledge or information for the research domain.
Motivation & Gap	MG	Highlights research motivations or identifies gaps in existing literature.
Comparison	CMP	Compares the current work or other baselines with specific prior methods.
Basis & Support	BS	Provides theoretical foundations or concepts supporting the current study.
Method & Data	MTD	Refers to specific methods, algorithms, tools, or datasets utilized in the research.
Validation	VLD	Provides empirical evidence, results, or evaluation metrics to validate claims.
Evolution	EV	Describes the historical evolution or development of a methodology.
Critique	CRT	Expresses critiques, conflicts, or opposing views regarding prior work.

Table 2: Detailed descriptions of the 8-category Citation Intent Taxonomy (CAP-8).

3.2 Citation Location Perception Data

We formulate citation location perception as a sequence restoration task. Given a paragraph with all original citations stripped, the task requires predicting a unified `[#CITE#]` marker at every legitimate citation position. Beyond position recovery, we use DeepSeek-V3 to annotate the necessity of each ground-truth marker, classifying it as either *Mandatory* (e.g., references to specific methods, algorithms, or datasets) or *Optional* (e.g., general domain background). By jointly supervising position and necessity, the dataset provides a more precise and context-aware location signal. To support this task, we construct 31,118 samples (comprising 28,006 for training and 3,112 for testing)

pairing the stripped input text with target outputs that mark every citation position together with its necessity label. A complete example is shown in Appendix C.1.

```
[Index 0]
<reasoning>
The purpose of this citation is to provide an
example of how testing can reveal biases or
spurious correlations learned by a model,
supporting the claim about the value of
invariance testing. </reasoning>
<keywords>
[‘‘spurious correlations’’, ‘‘text
classification’’, ‘‘counterfactuals’’,
‘‘robustness’’] </keywords>
```

Listing 1: An example of the synthesized training data for QueryPlanner.

3.3 Query Planning Data

Moving beyond citation location, the QueryPlanner dataset focuses on semantic intent. Its objective is to bridge the vocabulary gap between a paragraph's vague local context and the precise metadata of the target reference. Given a paragraph and a target citation position, the task is to generate both a reasoning chain that infers the citation intent and a keyword set that can retrieve the target reference. We isolate 7,079 high-density paragraphs (comprising 6,779 for training and 300 for testing) whose BibTeX entries contain complete titles and abstracts, and use DeepSeek-V3 in a posterior setting: with full access to the target reference's metadata, it retroactively deduces a logical reasoning path (<reasoning>) and an optimal keyword set (<keywords>) for each citation (Listing 1). This posterior view yields keywords that effectively retrieve the target.

To verify keyword quality, we run a cross-model consistency check on 100 random samples. We substitute the keywords in 50 of these samples with keywords from unrelated instances to form negatives, and five LLMs binary-classify whether the keywords align with the target reference. The 90.6% average accuracy (Cohen's $\kappa = 0.805$) confirms the reliability of the synthesized data. To analyze the logical distribution of training data, we run LLM-based motivation classification over 82,881 citation nodes; allowing secondary motivation labels expands the total to 93,403. This yields the 8-category Citation Intent Taxonomy (CAP-8), with category definitions in Table 2 and the corpus distribution in the intent panel of Figure 2.

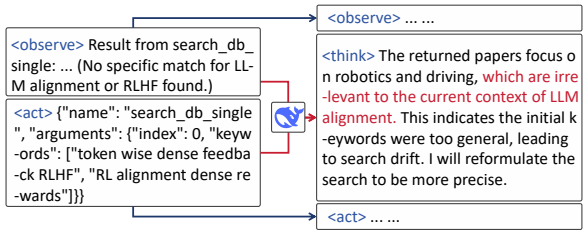

Figure 3: Data Generation Pipeline for Trajectory Thinking Process.

3.4 Reflective Trajectory Synthesis

The third subset trains Master Brain to orchestrate the workflow and recover from retrieval failures. Each sample is a multi-turn trajectory in an observe-think-act format: the agent observes the environment state (<observe>), articulates a plan (<think>), and issues an action (<act>) such as a tool call or a final citation choice. To synthesize supervision, we fix <observe> and <act> from successful executions and use DeepSeek-V3 to retroactively deduce the connecting <think> (Figure 3). We construct 2,000 trajectories in total: 500 "gold" paths and 1,500 reflection-enhanced variants.

For the gold paths, we inject real distractors from the local database so that the model must state within <think> why it prefers the target reference over each alternative. For the reflection-enhanced variants, we inject vague initial keywords to simulate retrieval drift (e.g., returning irrelevant survey papers); the model must diagnose the failure, refine keywords from broad domains to specific entities, and issue a secondary retrieval before succeeding. A complete example is in Appendix C.2.

To guarantee robust generalization and prevent data leakage, we construct a separate test set for our end-to-end citation prediction evaluation. Specifically, we randomly sample 50 paragraphs from CVPR'25, EMNLP'24, ICCV'23, and NeurIPS'25, yielding 200 paragraphs in total. This evaluation set is strictly disjoint from our training data pool and contains 1,023 unique target test papers. An audit confirms that none of the 2,640 training citation decisions involved any of the 1,023 test papers.

4 Method

Figure 4 illustrates the framework of ReCite, and we detail each module below.

4.1 Logical Breakpoint Perception

CiteLocator predicts citation boundaries and their necessity labels (Mandatory or Optional). We fine-

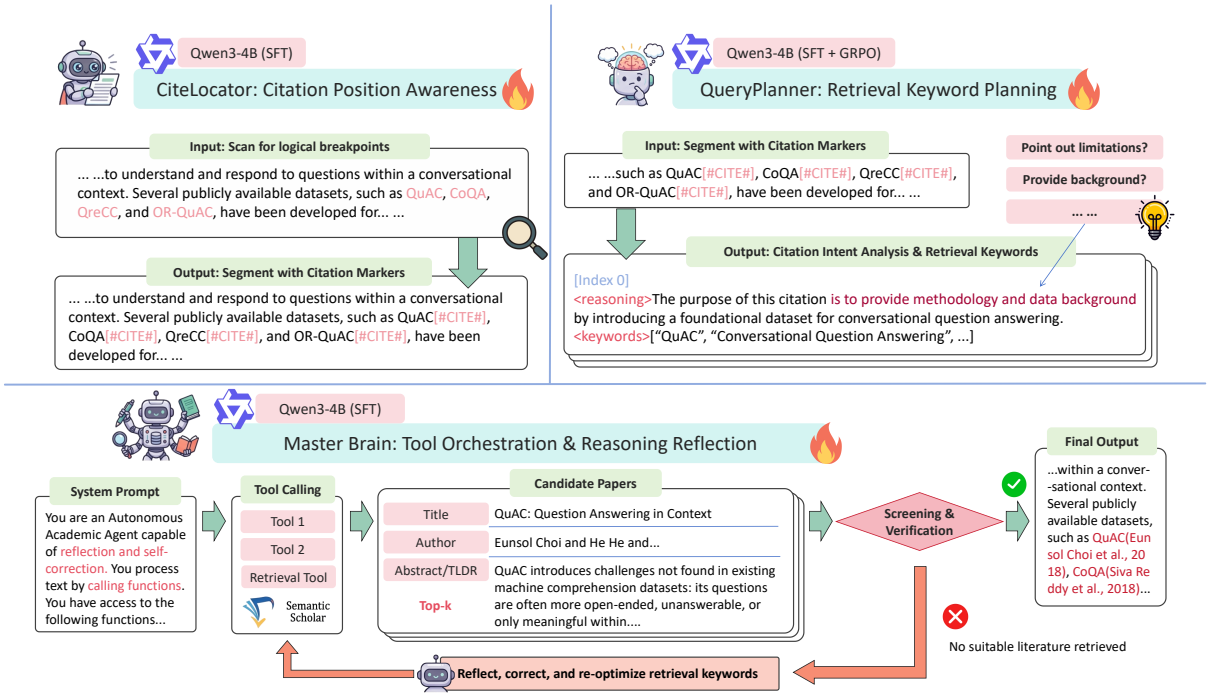

Figure 4: The overall architecture of our proposed **ReCite** framework. The system decouples the automatic citation process into three iterative stages: citation location perception (CiteLocator), intent-aware query planning (QueryPlanner), and task scheduling with reflective verification orchestrated by the Master Brain.

tune Qwen3-4B (Yang et al., 2025) via SFT on the location perception data, registering the unified `[#CITE#]` marker as a new tokenizer entry so that it is not fragmented during encoding. Because citation tokens are extremely sparse compared to regular text tokens, standard cross-entropy collapses to the majority class; we mitigate this with an asymmetric cross-entropy that places heavier loss weight on citation marker tokens, penalizing missed or hallucinated boundaries.

4.2 Intent-Aware Query Planning

QueryPlanner bridges the local context and precise retrieval metadata by generating a reasoning chain and a keyword set per citation position. We first fine-tune Qwen3-4B via SFT on the query planning data, then apply GRPO (Shao et al., 2024) to elicit deeper reasoning. For each query, GRPO samples a group of $G = 4$ candidate outputs and updates the policy with their relative advantages, eliminating the need for a separate value network. The composite reward combines three components: *format compliance*, which penalizes deviations from the required `<reasoning>`/`<keywords>` layout; *entity recall*, against the target reference; and *intent alignment*, against the ground-truth intent.

4.3 Task Orchestration

Master Brain is the central orchestrator of the agentic workflow. Like the other two modules, it is built on Qwen3-4B and trained via SFT on the reflective trajectory data to follow an observe-think-act state machine. At inference, Master Brain first invokes CiteLocator to identify all citation positions in the draft. For each position, it calls QueryPlanner to generate search keywords, queries the Semantic Scholar API to fetch candidate metadata (title, abstract, authors, year, citation count), and verifies whether the evidence supports the local claim. When no candidate provides consistent evidence, the agent enters a reflective re-retrieval loop: it emits a failure-attribution `<think>`, refines keywords from broad domains to specific entities, and issues a secondary retrieval. Because retrieval queries hit authentic databases, fabrication is impossible by construction; the residual challenge of misattribution is directly targeted by this verification and self-correction loop.

5 Experiments

5.1 Settings

Baselines The base model for all three modules is Qwen3-4B-Instruct (Yang et al., 2025). We compare our framework against three categories

Method	Size	Overall-Strictly (%)			Lenient Evaluation (%)			Position-Only (%)		
		P	R	F1	P	R	F1	P	R	F1
Models via zero-shot prompt pipeline										
GLM-4-Plus (GLM et al., 2024)	—	9.83	10.27	10.05	17.22	17.99	17.60	47.08	49.19	48.11
Mimo-V2.5-Pro (Xiaomi, 2026)	—	26.40	27.72	27.04	29.19	30.63	29.89	75.07	78.78	76.88
GPT-4o-mini (OpenAI et al., 2024)	—	7.20	8.65	7.86	12.73	15.29	13.89	44.89	53.90	48.98
GPT-5.1-Chat (Singh et al., 2025)	—	8.99	12.12	10.32	14.48	19.54	16.63	35.32	47.64	40.57
Kimi-k2-preview (Kimi et al., 2026)	—	19.06	21.24	20.09	24.32	27.10	25.64	55.86	62.24	58.88
Qwen3.6-Plus (Yang et al., 2025)	—	28.09	35.44	31.34	31.09	39.23	34.69	68.60	86.50	76.52
Qwen3.6-27B (Yang et al., 2025)	27B	22.89	29.62	25.83	25.58	33.08	28.85	62.34	80.48	70.26
DeepSeek-V4-Flash (DeepSeek-AI, 2026)	284B	31.23	35.21	33.10	34.52	38.89	36.57	72.95	82.11	77.26
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	1.6T	31.29	32.28	31.77	35.25	36.34	35.79	75.97	78.32	77.13
General Agents (End to End via Mimo-V2.5-Pro)										
Claude Code (Anthropic, 2026)	—	17.73	27.26	21.49	21.35	32.82	25.87	25.51	39.23	30.92
Hermes Agent (NousResearch, 2026)	—	22.39	21.24	21.80	26.38	25.02	25.68	28.01	26.56	27.27
OpenClaw (Steinberger, 2026)	—	18.09	7.03	10.12	25.25	9.81	14.13	29.42	11.43	16.46
Ours										
ReCite-Base	4B	36.92	35.75	36.33	43.46	42.08	42.76	89.80	86.11	87.53
ReCite-SFT	4B	37.48	37.45	37.47	47.14	47.07	47.10	89.80	**89.66**	**89.73**
ReCite-SFT(CAP-8)	4B	**39.22**	**39.07**	**39.15**	**55.27**	**55.02**	**55.14**	**89.92**	89.51	89.71

Table 3: End-to-end citation prediction evaluation results. We report Precision (P), Recall (R), and F1-score across three metrics: Strict End-to-End, Lenient Evaluation, and Position-Only. CAP-8 denotes the model fine-tuned with citation taxonomy information. Bold indicates the best result in each column.

Model	Size	Strategy	Mandatory (%)			Optional (%)			Overall (%)		
			P	R	F1	P	R	F1	P	R	F1
Baseline Models											
GPT-5.1-Chat (Singh et al., 2025)	—	Direct	22.01	59.72	32.17	9.21	58.50	15.92	27.73	59.39	37.81
GPT-4o-mini (OpenAI et al., 2024)	—	Direct	24.62	53.61	33.75	11.34	57.17	18.92	31.25	54.57	39.74
Kimi-k2-preview (Kimi et al., 2026)	—	Direct	31.27	61.05	41.36	14.36	61.28	23.27	38.38	61.11	47.15
MiniMax-M2.5 (MiniMax-AI, 2026)	—	Direct	33.52	56.37	42.04	14.71	52.52	22.98	40.35	55.34	46.67
GLM-4-flash (GLM et al., 2024)	—	Direct	25.81	16.73	20.30	12.07	17.96	14.43	32.67	17.06	22.41
Qwen3-Max (Yang et al., 2025)	—	Direct	32.87	28.90	30.75	15.41	29.28	20.19	40.18	29.00	33.69
Qwen3-Max (Yang et al., 2025)	—	CoT	34.43	28.42	31.14	16.00	28.09	20.39	41.71	28.33	33.74
DeepSeek-V3 (DeepSeek-AI et al., 2025)	671B	Direct	30.91	**62.66**	41.40	13.91	**61.65**	22.70	37.85	**62.39**	47.12
DeepSeek-V3 (DeepSeek-AI et al., 2025)	671B	CoT	29.23	49.09	36.64	12.41	45.88	19.54	35.68	48.23	41.02
Qwen3.5-27B (Yang et al., 2025)	27B	Direct	36.13	55.86	43.88	15.83	50.56	24.11	42.98	54.44	48.04
Qwen3-4B (Yang et al., 2025)	4B	Direct	13.98	9.70	11.45	5.55	9.56	7.03	18.12	9.66	12.60
Qwen3-4B (Yang et al., 2025)	4B	CoT	8.98	3.99	5.52	2.59	2.92	2.74	11.12	3.70	5.55
Ours											
CiteLocator	**4B**	**Direct**	**67.93**	60.61	**64.06**	**42.90**	58.57	**49.52**	**74.16**	60.06	**66.37**

Table 4: Citation location prediction results on the test set ($N = 3,112$). We report Precision (P), Recall (R), and F1-score for Mandatory, Optional, and Overall categories. Bold indicates the best result in each column.

of baselines: (i) zero-shot LLM pipelines that mirror our workflow; (ii) **General Agents** (Claude Code (Anthropic, 2026), Hermes Agent (NousResearch, 2026), and OpenClaw (Steinberger, 2026)) running end-to-end on a Mimo-V2.5-Pro inference engine; and (iii) heuristic and human baselines for location-only evaluation (GM-s2orc, GM-s2orc-H, and human experts (Buscaldi et al., 2024)). The full LLM baseline list is in Appendix A; detailed system prompts are in Appendices C.3-C.5.

Evaluation Metrics Metrics strictly align with our architecture. CiteLocator uses Precision (P), Recall (R), and F1-score for *Mandatory* and *Optional* citations. QueryPlanner uses Token F1 (R-F1) for reasoning, alongside Token F1/Recall (K-F1, K-TR), Phrase Recall (K-PR), Jaccard Similarity (K-Jac), and an LLM-as-a-judge mechanism for keyword quality. End-to-end workflows are evaluated at three strictness levels: *Overall-Strictly* (exact target match), *Lenient* (relevant background/baselines), and *Position-Only*. Details on the evaluation metric definitions and calculations are provided in Appendix B.

Parameters For SFT, we use an 8-bit Paged AdamW optimizer (learning rate 2×10^{-5}, 3 epochs, cosine decay, 5% warmup). Maximum sequence lengths are dynamically set: 1,024 for CiteLocator, 4,096 for QueryPlanner, and 9,216 for the Master Brain to preserve multi-turn trajectories. In the GRPO phase for QueryPlanner, the learning rate is 5×10^{-7} with a sample group size of $G = 4$.

Model	Size	Reasoning		Keyword Extraction				
		R-F1 (%)	R-Jdg	K-F1 (%)	K-TR (%)	K-PR (%)	K-Jac (%)	K-Jdg
Baseline Models								
GPT-5.1-Chat (Singh et al., 2025)	—	52.25	7.48	39.05	44.03	12.16	7.87	6.26
MiniMax-M2.5 (MiniMax-AI, 2026)	—	53.88	7.86	41.19	47.19	13.73	8.75	6.53
Qwen3-Max (Yang et al., 2025)	—	53.91	7.33	37.01	36.56	15.38	10.83	5.95
GLM-4-flash (GLM et al., 2024)	—	21.73	3.04	16.09	19.95	6.65	3.98	2.63
Kimi-k2-preview (Kimi et al., 2026)	—	54.12	7.75	45.80	**55.30**	**25.45**	**16.42**	**7.12**
DeepSeek-V3 (DeepSeek-AI et al., 2025)	671B	**58.63**	**7.88**	**46.25**	49.18	22.59	15.13	6.65
Qwen3.5-27B (Yang et al., 2025)	27B	50.55	7.56	40.99	43.62	16.73	11.14	6.33
Ours								
QueryPlanner(SFT)	4B	56.12	7.21	37.74	43.56	16.87	10.38	5.93
QueryPlanner(SFT+GRPO)	4B	57.81	7.64	39.06	42.47	17.85	11.00	6.44

Table 5: Comprehensive evaluation results about intent-aware query planning on the test set ($N = 300$). Bold indicates the best result in each column.

5.2 Main Results

As shown in Table 3, ReCite-SFT(CAP-8) achieves the highest Strict F1 of 39.15%, outperforming every baseline including DeepSeek-V4-Pro and Qwen3.6-Plus, demonstrating the effectiveness of our decoupled agentic design. This margin is especially notable despite the larger baselines likely benefiting from data leakage, since their pre-training corpora almost certainly cover our 2024-2025 test papers. In addition, adding CAP-8 over ReCite-SFT yields a clear gain on both Strict and Lenient F1, showing that fine-grained intent supervision substantially improves candidate verification. By comparison, General Agents (Claude Code, Hermes Agent, OpenClaw) fall well behind, indicating that black-box end-to-end agents cannot recover from retrieval drift, and that a task-specific agent design remains necessary in this setting.

While ReCite achieves state-of-the-art results, its Strict F1 (39.15%) is heavily constrained by the severity of exact-match evaluation, which requires perfect joint accuracy across sequential stages and is confounded by subjective author biases. Since scientific claims can often be validated by multiple appropriate papers, the 15.99% gain in Lenient F1 (55.14%) demonstrates ReCite's ability to discover logically supportive alternatives that are semantically valid but not explicitly cited by the original authors (a detailed qualitative case study is provided in Appendix D).

5.3 Module-Level Evaluation

Beyond end-to-end performance, we also evaluate CiteLocator and QueryPlanner individually against module-specific baselines.

CiteLocator. On location prediction (Table 4), CiteLocator outperforms all LLM baselines by a wide margin, reaching Overall F1 of 66.37% versus 48.04% for the next-best baseline (Qwen3.5-27B), with consistent gains across both Mandatory and Optional categories. On the out-of-domain BUSCALDI benchmark (Table 6), which categorizes tasks into coarse-grained (sentence-level necessity) and fine-grained (token-level placement) predictions, CiteLocator significantly narrows the gap to human experts on exact location prediction. By outperforming GM-s2orc-H on Fine-grained, it demonstrates that our learned rhetorical signals generalize effectively beyond the training corpus.

Type	Method	P (%)	R (%)	F1 (%)
Coarse-grained	Scientist	76.2	88.4	81.8
	GM-s2orc	78.2	78.2	78.2
	GM-s2orc-H	80.2	82.6	81.4
	CiteLocator	66.7	87.0	75.5
Fine-grained	Scientist	57.5	66.6	61.7
	GM-s2orc	44.9	44.9	44.9
	GM-s2orc-H	49.2	50.7	50.0
	CiteLocator	46.7	60.9	52.8

Table 6: Performance comparison with existing baselines on the citation prediction task. The "-H" suffix indicates methods utilizing extra heuristic rules.

QueryPlanner. On reasoning and keyword generation (Table 5), QueryPlanner-SFT alone reaches Token F1 of 56.12%, approaching the 58.63% of DeepSeek-V3 (671B) and clearly beating the 27B open-source baseline. Adding GRPO further lifts reasoning quality and most keyword metrics, with the composite reward (format + entity recall + intent alignment) pushing the model from surface copying to deeper entity grounding.

5.4 Ablation Study

To validate our decoupled framework, we individually replace CiteLocator, QueryPlanner, and the

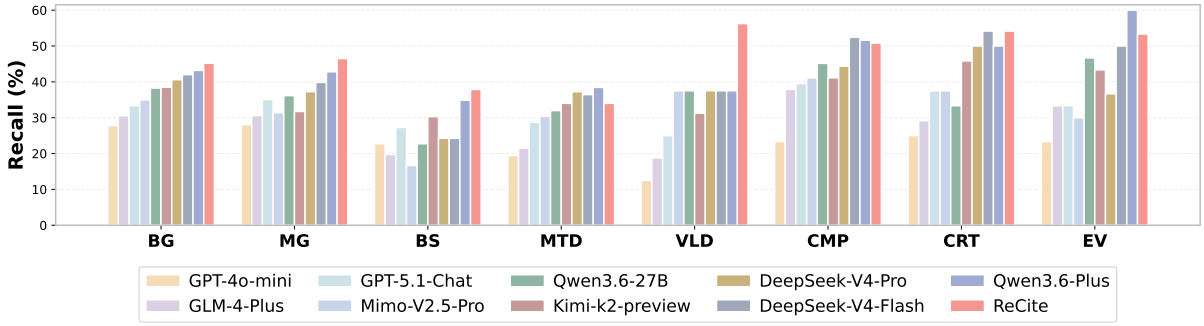

Figure 5: Performance comparison of various models across different citation intent categories. The evaluation is based on the Recall metric within the CAP-8 taxonomy.

Method	P (%)	R (%)	F1 (%)
Overall-Strictly			
ReCite	**37.48**	**37.45**	**37.47**
w/o CiteLocator	2.98 (-34.50)	2.55 (-34.90)	2.75 (-34.72)
w/o QueryPlanner	15.79 (-21.69)	15.75 (-21.70)	15.77 (-21.70)
w/o Master Brain	36.92 (-0.56)	35.75 (-1.70)	36.33 (-1.14)
Lenient Evaluation			
ReCite	**47.14**	**47.07**	**47.10**
w/o CiteLocator	7.05 (-40.09)	6.02 (-41.05)	6.49 (-40.61)
w/o QueryPlanner	30.88 (-16.26)	30.81 (-16.26)	30.85 (-16.25)
w/o Master Brain	43.46 (-3.68)	42.08 (-4.99)	42.76 (-4.34)

Table 7: Ablation study on different components.

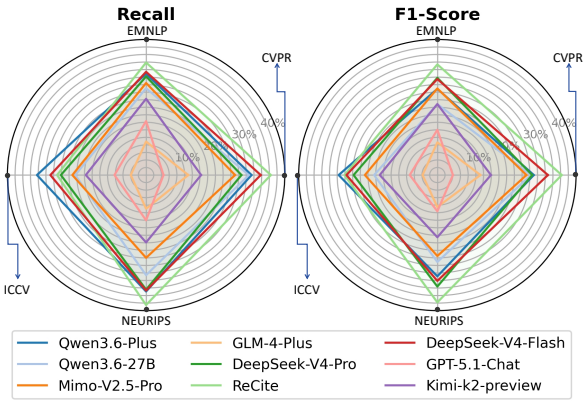

Figure 6: Performance comparison of different models based on the Recall and F1-Score.

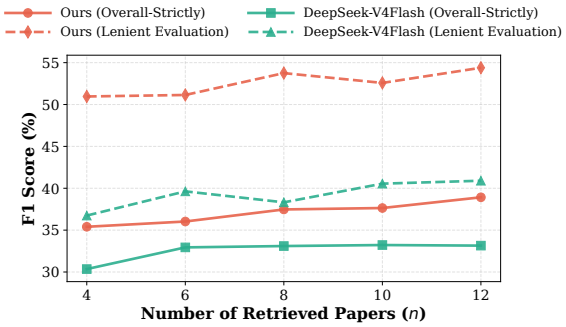

Figure 7: Performance variation of the proposed system and the baseline across different numbers of retrieved candidate papers (Top-k).

Master Brain with the Qwen3-4B-Instruct baseline (Table 7). Replacing CiteLocator causes a severe drop in Strict F1 to 2.75%, proving that logical breakpoint perception is the foundational prerequisite. Similarly, the absence of QueryPlanner lowers Strict F1 to 15.77%, highlighting the necessity of intent-aware query planning. Also, removing the Master Brain's reflective verification degrades the Lenient F1 from 47.10% to 42.76%, indicating its crucial role in filtering out noise. Ultimately, the complete system achieves optimal performance, demonstrating that these modules are mutually reinforcing and indispensable.

5.5 In-depth Analysis

Analysis of Cross-Venue Data Robustness Figure 6 breaks down end-to-end F1 by the source conference of each test paragraph. ReCite stays balanced across all four sub-domains, while several general-purpose baselines fluctuate or skew toward specific venues. This suggests that ReCite captures domain-general rhetorical signals rather than surface vocabulary overlap, supporting cross-domain transfer within computer science.

Analysis of Intent-Level Recall Figure 5 reports recall across the CAP-8 intent categories. ReCite achieves higher recall than most baselines across categories, indicating that explicit intent supervision helps the agent align retrieval with author motivation rather than relying on semantic overlap.

Analysis of Retrieval Candidate Quantity As k varies from 4 to 12 (Figure 7), the end-to-end F1 scores of both ReCite and the DeepSeek-V4-Flash baseline improve up to k = 8, after which further improvements are limited by context noise. More-

over, a larger k inflates input length and inference cost; we therefore fix $k = 8$ as the optimal trade-off between evidence sufficiency and latency.

6 Conclusion

In this paper, we introduced ReCite, a decoupled agentic reasoning framework for faithful citation. ReCite integrates citation location perception, intent-aware query planning, and reflective verification to address misattribution while eliminating fabricated references through retrieval from authentic academic databases. To train these capabilities, we constructed a large-scale reasoning-oriented dataset from LaTeX sources and synthesized supervision for location prediction, citation intent inference, keyword planning, and self-corrective trajectories. Extensive experiments show that ReCite achieves stronger strict citation accuracy than substantially larger generative models and general-purpose agents, demonstrating the effectiveness of our lightweight modules. Ablation and intent-level analyses confirm that each component contributes to robust citation grounding. By establishing citation as verifiable reasoning, ReCite provides a practical foundation for AI-assisted academic writing that is more faithful, accountable, and trustworthy.

Limitations

While **ReCite** demonstrates strong performance in specific citation tasks, several limitations remain. First, our specialized lightweight model inherently trails massive general-purpose LLMs in open-ended reasoning outside the targeted workflow. Second, the system currently operates exclusively on uni-modal text. Since real-world citations often rely on charts or pseudocode, future work will integrate Vision-Language Models (VLMs) to support multimodal verification. Third, because the framework was trained entirely on computer science papers, its cross-disciplinary generalization (e.g., to the humanities or biomedical fields) remains unexplored and requires systematic evaluation.

Ethical Considerations

Our research directly addresses a critical ethical issue in AI-assisted scientific writing: the fabrication of references, which severely misleads readers and undermines academic integrity. By enforcing a rigorous verification loop grounded in real-world academic databases, our framework mitigates the propagation of hallucinated knowledge.

We recognize the ethical concerns surrounding data privacy and intellectual property. Academic research often involves highly confidential, unpublished data, and relying on closed-source, cloud-based LLMs poses inherent compliance and data leakage risks. A core motivation for developing our framework on a lightweight 4B parameter model is to facilitate future model quantization and pruning. This paves the way for fully offline, on-device deployment, ensuring that researchers can leverage agentic citation assistance without compromising sensitive intellectual property.

Acknowledgements

We acknowledge the use of AI assistants during the writing and coding phases of this research. These tools were utilized solely to refine linguistic phrasing and assist with coding, while all scientific concepts, experimental designs, and final analyses were developed entirely by the authors.

References

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. Large language models reflect human citation patterns with a heightened citation bias. In *Findings of NAACL*, pages 6844–6879.

Anthropic. 2026. Claude code.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP*, pages 3615–3620.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222.

Davide Buscaldi, Danilo Dessí, Enrico Motta, Marco Murgia, Francesco Osborne, and Diego Reforgiato Recupero. 2024. Citation prediction by leveraging transformers and natural language processing heuristics. *Information Processing & Management*, 61(1):103583.

Ege Yiğit Çelik and Selma Tekir. 2025. CiteBART: Learning to generate citations for local citation recommendation. In *Proceedings of EMNLP*, pages 1703–1719.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT*, pages 3586–3596.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of ACL*, pages 2270–2282.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, and 1 others. 2018. Science of science. *Science*, 359(6379):eaao0185.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, pages 6465–6488.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Nianlong Gu, Yingqiang Gao, and Richard HR Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *European conference on information retrieval*, pages 274–288. Springer.

Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2014. Refseer: A citation recommendation system. *IEEE/ACM Joint Conference on Digital Libraries*, pages 371–374.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Team Kimi, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Bobo Li, Rui Wu, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee, and Wynne Hsu. 2026a. Taming actor-observer asymmetry in agents via dialectical alignment. In *Proceedings of ACL*, pages 24068–24084.

Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2026b. Step-grpo: Enhancing reasoning quality and efficiency via structured prm-based reinforcement learning. In *Proceedings of AAAI*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of ACL*.

Zoran Medić and Jan Snajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 97–103.

MiniMax-AI. 2026. Minimax m2.5 large language model.

NousResearch. 2026. Hermes agent documentation.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Yu Pei, Yongping Du, and Xingnan Jin. 2025. Fover: First-order logic verification for natural language reasoning. *Trans. Assoc. Comput. Linguistics*, 13:1340–1359.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. *CoRR*, abs/2601.18724.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in NeurIPS*, pages 8634–8652.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry,

Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Peter Steinberger. 2026. Openclaw.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. 2025. Scholarcopilot: Training large language models for academic writing with accurate citations. *Preprint*, arXiv:2504.00824.

Xiaomi. 2026. Mimo-v2.5.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. ALiiCE: Evaluating positional fine-grained citation generation. In *Proceedings of NAACL-HLT*, pages 545–561.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Yao Zhang, Yude Bai, Minhong Dong, Keqing Cen, Ji Zhang, Qiang Hu, Wei Ma, Yongqiang Lyu, Ruitao Feng, Xiaohong Li, Junjie Wang, Lingxiao Jiang, and Yang Liu. 2026. Limr: Intent-aware mashup api recommendation via llm-augmented multi-scale fusion. *IEEE Transactions on Services Computing*.

A Experiment Settings

Zero-shot LLM Baselines. For end-to-end and module-level evaluation we benchmark against proprietary models (GPT-4o-mini (OpenAI et al., 2024), GPT-5.1-Chat (Singh et al., 2025), Kimi-k2-preview (Kimi et al., 2026), GLM-4-Plus, GLM-4-Flash (GLM et al., 2024), Qwen3.6-Plus, Qwen3-Max (Yang et al., 2025), Mimo-V2.5-Pro (Xiaomi, 2026), MiniMax-M2.5 (MiniMax-AI, 2026)) and open-source LLMs (Qwen3.5-27B, Qwen3-4B (Yang et al., 2025), DeepSeek-V3 (DeepSeek-AI et al., 2025), DeepSeek-V4-Flash, DeepSeek-V4-Pro (DeepSeek-AI, 2026)), all evaluated zero-shot under a prompt pipeline mirroring our workflow.

General Agents. The three end-to-end agents (Claude Code, Hermes Agent, OpenClaw) receive a single end-to-end citation objective and orchestrate their own tool calls; for fair comparison, all three run on Mimo-V2.5-Pro.

Heuristic and Human Baselines. For location-only evaluation, we compare against GM-s2orc, a GPT-2 based generative model fine-tuned on the s2orc dataset, and its variant GM-s2orc-H, which incorporates additional post-hoc NLP heuristics. Scientist denotes human annotator performance from (Buscaldi et al., 2024). To evaluate out-of-domain generalization, we utilize their introduced gold standard dataset, which comprises 133 sentences randomly sampled from recent arXiv papers and strictly annotated by three senior researchers.

Implementation Details. We train and evaluate ReCite on NVIDIA RTX 4090 D and A800 GPUs, with DeepSpeed for distributed training and vLLM (PagedAttention) for high-throughput inference.

B Details of Evaluation Metrics

We detail the evaluation metrics, strictly aligning with the three modules of the ReCite framework.

B.1 Main Results

The end-to-end evaluation assesses the final citation mapping at three strictness levels:

- **Overall-Strictly:** Requires both correct identification of the citation position AND successful retrieval of the exact ground-truth paper.
- **Lenient Evaluation:** Accepts highly relevant substitute papers, acknowledging that multiple papers can logically support a claim. This is assessed via an **LLM-as-a-Judge** that outputs a binary score (1 for academically appropriate, 0 otherwise) based on local context and candidate metadata.
- **Position-Only:** Evaluates structural perception alone. A prediction is correct if it successfully identifies the logical breakpoint, regardless of the retrieved candidate.

B.2 Citation Location Perception (CiteLocator)

We evaluate the generated `[#CITE#]` placeholders against ground-truth breakpoints using standard Precision (P), Recall (R), and F1-score ($F1$):

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{1}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{2}$$

These metrics are computed separately for *Mandatory* (specific evidentiary backing) and *Optional* (general background) citations.

B.3 Intent-Aware Query Planning (QueryPlanner)

To evaluate the QueryPlanner, we utilize the following metrics:

- **Token Metrics (R-F1, K-F1, K-TR):** Unigram F1 overlap for the generated `<reasoning>` (R-F1) and `<keywords>` (K-F1), along with the keyword Token Recall (K-TR), evaluated against the ground-truth.
- **Phrase Recall (K-PR):** The proportion of ground-truth core entities successfully extracted in the `<keywords>` list.
- **Jaccard Similarity (K-Jac):** The intersection over union of the predicted (K_{pred}) and ground-truth (K_{gt}) keyword sets:

$$\text{K-Jac} = \frac{|K_{pred} \cap K_{gt}|}{|K_{pred} \cup K_{gt}|} \tag{3}$$

- **LLM-as-a-Judge:** An advanced LLM scores the keyword quality (0-10) based on logical soundness and retrieval feasibility.

C Data Examples and System Prompts

C.1 Examples of Location Perception Data

Listing 2: Example input and output for citation location perception data.

```
input: Most state-of-the-art FSRL methods adopt
a meta-learning framework, where a base model
is pre-trained... they predominantly rely on KG
embedding models like TransE or DistMult to
exploit relational information.

output: Most state-of-the-art FSRL methods
adopt a meta-learning framework [#CITE#], where
a base model is pre-trained... they
predominantly rely on KG embedding models like
TransE [#CITE#] or DistMult [#CITE#] to exploit
relational information.
```

C.2 Example of Reflective Trajectory Data

Listing 3: Example of multi-turn reflective trajectory data synthesis, demonstrating self-correction.

```
=== Turn 3 Assistant ===
<think>
The initial intent analysis suggested keywords
such as reinforcement learning feedback''. I
will use these to search for candidate papers
for index 0.
</think>
<act>
{name'': search_db_single'',arguments'':
{index'': 0,keywords'': [''reinforcement
learning feedback'']}}
</act>

=== Turn 4 User ===
<observe>
Result from search_db_single: ... (No specific
match for LLM alignment or RLHF found.)
</observe>

=== Turn 5 Assistant ===
<think>
The returned papers focus on robotics and
driving, which are irrelevant to the current
context of LLM alignment. This indicates the
initial keywords were too general, leading to
search drift. I will reformulate the search to
be more precise.
</think>
<act>
{name'':search_db_single'', arguments'':
{index'': 0, keywords'': [token-wise dense
feedback RLHF'', ''RL alignment dense
rewards'']}}
</act>
```

C.3 System Prompts for Pipeline Baselines

To ensure reproducibility and fair comparison, we provide the exact zero-shot system prompts used for our three-stage pipeline baselines.

Listing 4: Zero-shot Prompt for Citation Location Perception (CiteLocator Pipeline).

```
[System Prompt]
You are a professional academic copy-editor.

[User Prompt]
Task: Act as a professional academic
copy-editor. Insert the specified marker
[#CITE#] at places that require citations, such
as when mentioning prior works, existing
methods, algorithms, models, datasets or other
places where citations are needed.
Rules: Do not alter any original content. Do
not add extra preamble. Output only the
processed text.

Input Text:{text}
```

Listing 5: Zero-shot Prompt for Intent-Aware Query Planning (QueryPlanner Pipeline).

```
You are an expert academic citation analyst.

### Task
Analyze the context of the given text. For each
citation marker [#CITE#], infer what keywords
should be used to search for the paper that
needs to be cited. Each [#CITE#] marker in the
input must correspond to one [Index n] in the
output in order.

### Output Components
1. <reasoning>: A 1-2 sentence logical analysis
of the context explaining the core role of the
citation. MUST start with the fixed phrase:
''The purpose of this citation is to...''
2. <keywords>: 3-5 precise keywords to retrieve
the target document.

### Constraints
- Sequential Indexing: Start with [Index 0] and
increment by 1 for each marker.
- Output ONLY the structured content. Strictly
NO conversational filler.

### Output Format
[Index X]
<reasoning>...</reasoning>
<keywords>[''keyword 1'', ''keyword 2'',
...]</keywords>
```

Listing 6: Zero-shot Prompt for Candidate Verification (Search and Submit Pipeline).

```
[System Prompt]
You are an academic paper selector. Your goal
is to match the exact foundational paper needed
for the specific citation context.

[User Prompt]
You need to select the best citation for the
marker [Index {index}] (which corresponds to
the {index+1}-th [#CITE#] marker in the text).

ORIGINAL CONTEXT:
{marked_text}

Your reasoning for this citation was:
{reasoning}
```

```
Here are the candidate papers from the database:
{candidate papers}

Evaluate carefully based on the original
context. If you find the correct foundational
paper, reply with ONLY its ID. If NONE of these
papers match the context, reply with exactly
''NONE''.
```

C.4 System Prompts for General Agents

Listing 7: System Prompt for Autonomous General Agents (End-to-End via Mimo-V2.5-Pro).

```
You are an automatic citation assistant. Your
capability is to identify places in given text
paragraphs that require academic citations,
autonomously construct keywords to retrieve
real academic papers, and strictly follow the
rules to add standard citations. It is strictly
forbidden to search the original source article
to copy existing real citations, or to generate
fake papers.

STRICT CITATION RULES:
Citation Judgment: Identify positions that
require citations, such as when mentioning
prior works, existing methods, algorithms,
models, datasets, or other places where
citations are needed.
Literature Retrieval: Actively use keywords to
retrieve real literature for matching citations.
Citation Placement Rule: Insert exactly one
ordered citation mark (e.g., [1], [2], [3]...)
for each independent sentence or academic
claim. Multiple citations stacking at the same
position are strictly prohibited.
Bibliography Format: Append an exact line
titled ''Bibliography:'' at the very end of the
modified full text. Each bibliography entry
follows the fixed format: [Number] First
Author, ''Exact Full Paper Title''
No Extra Content: Do not add any redundant
explanations, comments, or modification notes
in the output text.

FIXED OUTPUT FORMAT (MANDATORY):
Output a single-line valid JSON object for each
processed entry, containing only two fields
with no extra content:
{''text'':''Full modified text with inserted
citations and final Bibliography section'',
''bibliography'':[''First Author, \''Exact Full
Paper Title\'']}

HARD CONSTRAINTS:
Never modify the original input file in any way.
All cited literature must be real and
retrievable; fake literature and copying
original citations are strictly prohibited.
Strictly follow the ''one sentence, one
citation'' rule; no citation stacking or
redundant citation marking.
```

C.5 System Prompt for Master Brain Agent

Listing 8: System Prompt for the Master Brain Agent (Task Orchestration and Reflective Verification).

```
You are an Autonomous Academic Agent capable of
reflection and self-correction. You process
text by calling functions. You have access to
the following functions: {functions}

CRITICAL RULES:
1. Sequential Workflow: mark_citations ->
analyze_intents -> (Loop starts) ->
search_db_single (Index 0) ->
evaluate_and_submit_single (Index 0) ->
search_db_single (Index 1) ... ->
finalize_chunk.
2. You MUST process each index one by one. Do
NOT search for Index 1 until you have submitted
Index 0.
3. Once all indices reported by analyze_intents
are submitted, call finalize_chunk.

OUTPUT FORMAT:
Every response MUST consist of two parts:
1. A reasoning process wrapped in
<think>...</think> tags.
2. A function call wrapped in <act>...</act>
tags.
```

D Case Study on the Multi-Option Reality in Citations

To illustrate the gap between Strict and Lenient metrics, we analyze a representative case from our test set. Consider the context: *"Compared with the global post-processing optimization used in previous work* [#CITE#]*, our method learns the descent direction..."* While the original authors cited GLAMR (Yuan et al., 2022), ReCite retrieved HuMoR (Rempe et al., 2021), a prominent framework that also utilizes a gradient-based global optimization loop for 3D human motion estimation. Both papers are academically appropriate and provide equivalent logical validation for the claim.

Under the Strict exact-match metric, retrieving HuMoR is penalized with a score of zero due to the string mismatch with the authors' specific bibliographic key. In contrast, our Lenient Evaluation correctly identifies HuMoR as a valid semantic alternative. This case highlights how Strict F1 significantly underrepresents ReCite's true citation faithfulness by ignoring acceptable multi-option academic realities and subjective author preferences.