RW-CITE: A LARGE-SCALE MULTI-DOMAIN BENCHMARK FOR CITATION RECOMMENDATION IN RELATED WORK GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable related-work generation begins with identifying the prior studies that establish a manuscript's context, differentiate its contributions, and support its claims. In the cold-start setting, however, a system must recover these references from a large and evolving literature using only the manuscript's title and abstract, without citation contexts or a predefined candidate set. We introduce **RW-Cite**, a multi-domain benchmark and end-to-end framework for evaluating this complete literature-discovery and citation-ranking process. RW-Cite applies multi-level topic retrieval to 3,144,903 arXiv paper metadata records and constructs domain-specific citation graphs with contextual evidence. From this corpus, we evaluate ten scientific domains spanning robotics, autonomous driving, biomedical imaging, astronomy, cosmology, condensed-matter physics, quantum computing, and scientific machine learning. The ten independently constructed graphs contain 257,393 node entries and 940,141 section-aware citation-edge entries. The RW-Cite framework combines learned structural shortlisting, a citation-aware SciBERT cross-encoder, a graph attention network (GAT), and fixed score-and-rank fusion. Across the evaluated domains, RW-Cite establishes strong citation-recommendation performance; on the EWM comparison benchmark, it attains $2.31\times$ the Hits@10 score of the strongest evaluated frontier-LLM-based pipeline. In a preliminary analysis of 640 training queries, an LLM assessed the 30 candidates ranked highest by RW-Cite for each query, yielding 19,200 query–candidate judgments; 96.2% were judged by the LLM evaluator as core or related prior work, providing complementary evidence of recommendation relevance beyond observed citation recovery. To support transparent and reproducible research, we publicly provide the benchmark datasets, temporal splits, graph and candidate manifests, together with the construction, framework, and evaluation code.

The source code is available at `https://anonymous.4open.science/r/RW-Cite/`; the datasets are available at `https://anonymous-hf.com/a/t71z3mpp2u2i/`.

1 INTRODUCTION

Recent LLM-based research agents can formulate hypotheses, run experiments, and draft scientific manuscripts (Lu et al., 2026; Gottweis et al., 2026), but these capabilities depend on reliable access to prior work. At the earliest stage of research, only a proposed title and abstract may be available; no citation context, reference list, or known citation neighborhood identifies where to search. This cold-start setting therefore requires a system to discover and rank supporting literature before novelty assessment, experimental design, and related-work generation can be grounded in evidence (Färber & Jatowt, 2020; Bhagavatula et al., 2018).

Existing evaluations often isolate reranking within a supplied candidate pool, obscuring whether relevant papers were absent from the corpus or lost during retrieval. They also vary in query context, gold construction, and temporal controls, making end-to-end comparisons difficult. A useful benchmark must connect domain construction, candidate coverage, and final ordering under one auditable contract while preventing query citations from becoming inference features.

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

We introduce **RW-Cite**, a large-scale multi-domain benchmark constructed automatically from 3,144,903 arXiv paper metadata records. Section-aware citation extraction, cited-paper closure, contextual evidence, frozen query splits, and candidate manifests produce ten independently constructed domain graphs containing 257,393 node entries and 940,141 citation-edge entries. RW-Cite separately measures candidate coverage and ranking quality, allowing failures to be localized to corpus construction, retrieval, shortlisting, or final ranking.

We also provide an evidence-aware reference framework. It learns a structural Top-400 shortlist, scores candidates with a citation-aware SciBERT cross-encoder (Beltagy et al., 2019; Nogueira & Cho, 2019) and graph attention (Veličković et al., 2018), and combines the channels through fixed score-and-rank fusion. Across ten domains, fusion consistently improves over either channel alone; on EWM, end-to-end comparisons demonstrate an advantage over the evaluated dense-retrieval and HLM-based pipelines. A complementary relevance analysis suggests that the recommendations extend beyond citations observed in the source bibliography.

Our contributions are:

- We formulate cold-start related-work recommendation from only a manuscript title and abstract and introduce an auditable evaluation contract that separates candidate coverage from final ranking.
- We automatically construct and publicly provide ten supervised domain benchmarks from 3,144,903 arXiv paper metadata records, with section-aware citation labels, contextual evidence, frozen splits, and reproducible manifests.
- We develop a reference framework combining structural shortlisting, citation-aware cross-encoding, graph attention, and score-and-rank fusion. Evaluation across ten domains shows consistent gains from combining textual and graph evidence, while end-to-end EWM comparisons demonstrate advantages over the evaluated dense-retrieval and HLM-based baselines.

2 Related Work

Citation recommendation is a complex multilayer problem that encompasses tasks with substantially different information requirements. Färber and Jatowt distinguish *local* recommendation, where a model observes a citation placeholder and its surrounding passage, from *global* recommendation, where references are proposed from a larger portion of a manuscript (Färber & Jatowt, 2020). This distinction affects candidate construction, supervision, and evaluation: a local model identifies a paper supporting a known citation context, whereas a global model must also infer which topics and scholarly relationships require citations. Our setting follows global, cold-start recommendation, where the system receives only the title and abstract of a new paper and returns a fixed set of references intended for its Introduction and Related Work. Prior content-based systems retrieve and rerank from manuscript text (Bhagavatula et al., 2018), while HLM-Cite adds LLM reasoning after embedding retrieval (Hao et al., 2024). MasterSet constructs graded must-cite labels with an LLM judge and identifies SciBERT–NTX as a strong dense baseline (Ratul et al., 2026); SciNCL and LitSearch likewise demonstrate the value of citation-informed representation learning and staged retrieval (Ostendorff et al., 2022; Ajith et al., 2024). RW-Cite adopts the stricter title–abstract interface while evaluating candidate discovery and ranking together rather than assuming a supplied pool.

Scholarly resources such as S2ORC and LitBench connect full text, resolved citations, and domain literature graphs (Lo et al., 2020; Varvarigos et al., 2026). ASPIRE and CORWA further show that co-citation structure and citation discourse encode relationships beyond topical similarity (Mysore et al., 2022; Li et al., 2022). RW-Cite unifies these signals in a task-aligned benchmark: cited-paper closure preserves recoverable references, section-aware extraction supplies labels and candidate evidence, and fixed splits and manifests make coverage and ranking jointly auditable. Existing resources address individual stages, but not this complete cold-start, multi-domain evaluation contract.

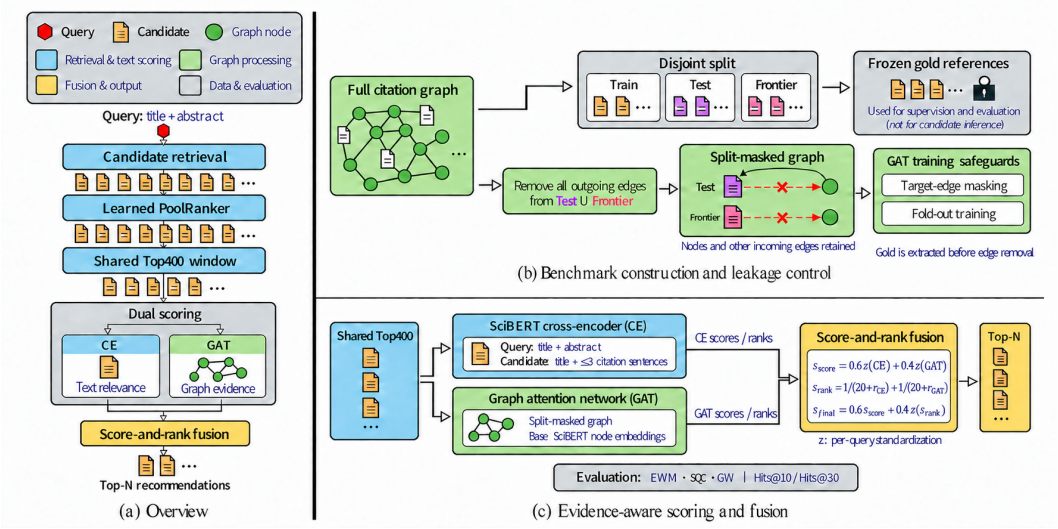

Figure 1: Overview of the RW-Cite framework. Given a query title and abstract, candidate retrieval and a learned structural ranker construct a shared Top-400 window. A SciBERT cross-encoder and graph-attention network independently score the candidates, after which score-and-rank fusion produces the final Top-10 recommendations. Gold references are extracted from the complete citation graph, whereas model construction and scoring use the split-masked graph obtained by removing outgoing edges from test and frontier sources.

3 BENCHMARK

RW-Cite converts a shared corpus of 3,144,903 arXiv paper metadata records into independently constructed, domain-specific citation-recommendation benchmarks. Each benchmark fixes its graph, contextual evidence, query split, candidate manifests, and gold references under a common data contract. Figure 1 summarizes both benchmark construction and the RW-Cite framework evaluated end to end.

3.1 TASK DEFINITION AND EVALUATION SCOPE

RW-Cite studies global citation recommendation in the cold-start setting. The input is limited to the title and abstract of a query paper; no manuscript body, citation placeholder, local citation context, reference list, or known citation neighborhood is available. A system must search the corresponding domain corpus and return an ordered list of ten papers suitable for grounding the query's Introduction and Related Work. This interface reflects the point at which a research agent must establish context and identify prior contributions before a complete manuscript or bibliography exists.

For each query, the benchmark derives a frozen gold set from resolved citations observed in its Introduction or Related Work-like sections. These citations provide scalable supervision from scholarly documents, but they are not treated as an exhaustive human judgment of every paper that could be useful. Observed citations may also reflect paper visibility and field-specific citation practices in addition to relevance (MacRoberts & MacRoberts, 1989; Nicolaisen, 2007). Accordingly, the primary evaluation measures recovery of the observed bibliography, while the separate relevance assessment in Section 5.4 examines potentially useful non-gold recommendations. Gold sets are not truncated to ten papers: Hits@10 and Hits@30 count recovered references, whereas recall normalizes by the full frozen set. Formal definitions appear in Appendix A.1.

The benchmark evaluates two failure modes separately. Candidate coverage asks whether a relevant paper survives corpus retrieval and structural shortlisting; ranking quality asks whether surviving papers are ordered near the head of the final list. This distinction is essential for end-to-end literature discovery: a reranker cannot recover a paper absent from its candidate pool, while high candidate re-

Table 1: Automatically constructed RW-Cite domain benchmarks. Retrieved papers are the approximate number of source directories retained before cited-paper closure; nodes and edges describe the frozen graph manifests paired with the reported ranking results. Counts are domain entries and need not be unique across domains.

Domain	Scientific area	Retrieved papers*	Graph nodes	Citation edges	Train queries	Test queries
EWM	Embodied AI	4,934	27,705	141,668	3,423	381
GW	Gravitational-wave physics	5,568	29,027	130,359	2,528	281
Driving	Autonomous driving	4,385	26,952	109,473	2,946	328
Cosmo	Physical cosmology	4,993	36,575	132,656	2,598	289
Radio	Radio astronomy	7,172	24,165	65,214	1,460	163
WSI	Computational pathology	4,972	19,547	50,660	1,665	185
Exo	Exoplanet astronomy	4,995	17,637	67,385	1,749	195
SQC	Superconducting quantum computing	7,891	20,470	87,273	2,636	293
FNO	Neural-operator turbulence	4,978	21,050	58,375	1,962	219
SCE	Strongly correlated electrons	9,008	34,265	97,078	1,832	204
Total	Ten domain graphs	∼58,896	257,393	940,141	22,799	2,538

call does not guarantee a useful Top-10. RW-Cite therefore records intermediate candidate manifests in addition to final rankings.

3.2 Automated Domain Construction

RW-Cite constructs each benchmark from the same 3,144,903-record arXiv metadata corpus through a domain-independent pipeline. First, each paper's three-level topic hierarchy is encoded with BGE (Xiao et al., 2024), and the resulting representation is matched against multiple domain prototypes. Rank and similarity filters produce a deduplicated seed corpus while allowing each domain definition to express several related research themes. This stage supplies broad topical coverage without using evaluation citations.

Second, the pipeline obtains source documents for retained seeds, identifies Introduction and Related Work-like sections, normalizes citation commands, and resolves bibliography entries to canonical paper identifiers. Third, *cited-paper closure* inserts a resolved cited paper even when it was not independently retrieved as a seed. Closure prevents foundational or older references from disappearing solely because they fall outside the recent topic-retrieval boundary. Finally, citation contexts are cleaned to remove comments, layout commands, figure fragments, malformed prefixes, and excessive citation-key lists without modifying graph topology. Query admission, split construction, and serialization then freeze the graph, labels, contextual evidence, and provenance.

This process derives supervision directly from observed citation behavior: after the domain prototypes and construction configuration are specified, no manual relevance labeling is required. The same pipeline can consequently be applied to additional scientific fields while preserving the benchmark schema and evaluation contract. Detailed extraction and cleaning rules appear in Appendix A.

3.3 Multi-Domain Benchmark Suite

Table 1 summarizes the ten benchmarks that we publicly provide. Together, the frozen graph manifests associated with the reported results contain 257,393 graph-node entries and 940,141 section-aware citation-edge entries, with 22,799 training queries and 2,538 test queries. Individual graphs range from 17,637 to 36,575 nodes and from 50,660 to 141,668 section-aware citation edges. The suite spans embodied and autonomous systems, astronomy and cosmology, medical imaging, quantum technologies, turbulence modeling, and condensed-matter physics. This variation tests whether a common recommendation contract remains useful across fields with different corpus sizes and citation structures.

The counts are domain entries rather than globally unique papers because the same work may legitimately participate in several scientific graphs. This design preserves domain-specific neighborhoods, evidence, and candidate spaces instead of collapsing the collection into a single heterogeneous graph. RW-Cite therefore contributes both a ten-domain resource and a reproducible procedure for producing further benchmark instances.

3.4 Data Schema and Automated Supervision

Each benchmark is a directed, section-aware citation graph. A node stores a canonical identifier, title, abstract when available, publication metadata, and source provenance. An edge (u, v) records that paper u cites paper v in an admitted section and retains the section label together with raw and cleaned citation-context text. The graph thus aligns four resources that are often distributed across separate datasets: a searchable domain corpus, reference-selection labels, candidate-side textual evidence, and structural relationships for graph-aware modeling.

RW-Cite distinguishes the complete extraction graph from the graph exposed to models. In the complete graph, a query's outgoing edges define its frozen gold references. We use these observed citations as reproducible gold labels. They represent the references selected by the original authors but do not exhaust all papers that may be relevant to the query. At inference, those edges and their citation sentences are hidden. Incoming citations from admissible source papers may still describe a candidate and therefore serve as candidate-side evidence. This asymmetry is intentional: the benchmark asks a model to infer the query's bibliography from previously available descriptions of candidate papers, not to reconstruct an already observed reference list. The exported text-view manifest records which evidence sentences were selected, their provenance, tokenization and truncation status, and whether an abstract fallback was required.

3.5 Temporal Splits and Leakage Controls

Eligible query sources must contain at least ten resolved outgoing citations, meet the required citation-context coverage, and fall within the configured gold-outdegree range. Admitted queries are ordered by publication time; the newest portion forms the test split and a one-month margin is retained as a corpus-only frontier. Training and test identifiers are disjoint, and test queries are excluded from structural, cross-encoder, and GAT training and from hyperparameter selection.

Before any model is constructed, outgoing edges from test and frontier sources are removed to form the split-masked graph. Their citations remain fixed as evaluation labels but cannot be used for retrieval, candidate evidence, or message passing. During optimization, all known gold identifiers for a query are excluded from its negative set, including positives not sampled in the current list; self-links and duplicate candidates are also removed. GAT training adds fold-specific masks and target-edge removal so that a supervised edge cannot reappear as a structural feature.

Every benchmark instance records train and test query-ID hashes, graph and source-list checksums, candidate-manifest fingerprints, publication metadata, and the applied graph mask. These artifacts make leakage controls independently auditable rather than implicit in a data-loading implementation. The complete audit fields are listed in Appendix A.3.

3.6 Evaluation Contract and Public Artifacts

All methods compared within a setting use the same query list, eligible corpus, full-gold labels, metric implementation, and output contract. End-to-end systems retain their native candidate-generation mechanisms because candidate discovery is part of the task. RW-Cite reports candidate coverage at intermediate depths, Hits@10 and Hits@30 for absolute recovery, nDCG@10 and MRR@10 for head ordering, and Recall@30 and Recall@100 for normalized deeper recovery. Per-query rankings and gold flags are retained so that every aggregate can be recomputed.

We publicly provide the ten domain datasets, temporal splits, graph and candidate manifests, contextual evidence with provenance, evaluation outputs, and the construction and scoring code. Each result is associated with its metadata snapshot, graph checksum, query split, configuration, and random seed. This release contract supports both standardized comparison and diagnosis of whether a missed reference was lost during corpus construction, candidate generation, shortlisting, or final ranking.

4 Reference Framework

RW-Cite provides a coarse-to-fine reference implementation for the benchmark. Given a query title and abstract, it constructs a high-recall universe, learns a structural Top-400 shortlist, scores that

shared window with complementary textual and graph models, and returns ten references through fixed score-and-rank fusion. The pipeline and hyperparameters are shared across domains, while model parameters are learned independently for each graph.

4.1 Candidate Universe and Structural Shortlisting

Scoring every graph node with a cross-encoder is impractical. RW-Cite first forms U_q as the deduplicated union of multi-view BGE retrieval, pseudo- relevance feedback from highly ranked titles, and citation-neighborhood expansion around query-relevant topic hubs. A weak second hop adds candidates supported by several focused hubs, including papers that may be missed by surface similarity. The query itself and its outgoing citation edges are excluded throughout.

Each $c \in U_q$ is represented by ten inexpensive features: graph-support count and log-support, lexical similarity, co-citation support, log in- and out-degree, dense-retrieval membership and rank bonus, a graph-support indicator, and embedding similarity. After standardization on training data, a logistic model estimates

$$p_{\text{struct}}(q, c) = \sigma\big(b + \mathbf{w}^\top \boldsymbol{\phi}(q, c)\big)\,. \tag{1}$$

The highest-scoring 400 candidates form W_q, the shared window used by both neural rankers. This learned reduction is evaluated separately from final ranking so that coverage gains cannot be mistaken for reranking gains. Full universe and support definitions appear in Appendix B.

4.2 Citation-Aware Cross-Encoder and Graph Ranker

The cross-encoder receives $x_q = [t_q; a_q]$ and a candidate representation $x_c = [t_c; \xi_1^c; \xi_2^c; \xi_3^c]$ containing up to three cleaned incoming citation sentences; a short abstract fragment is used when no context is available. SciBERT (Beltagy et al., 2019; Nogueira & Cho, 2019) jointly encodes (x_q, x_c) to produce $s_{\text{CE}}(q, c)$. For a sampled list L_q and included positives P'_q, it minimizes the multi-positive listwise loss

$$\mathcal{L}_q = -\frac{1}{|P'_q|} \sum_{c \in P'_q} \log \frac{\exp s(q, c)}{\sum_{c' \in L_q} \exp s(q, c')}. \tag{2}$$

In parallel, a two-layer GAT (Veličković et al., 2018) operates on SciBERT node embeddings and sparse incoming-citation, outgoing-citation, and co-citation relations. Relation-specific messages are combined with semantic similarity, query/candidate representations, and structural features to yield $s_{\text{GAT}}(q, c)$ over the same W_q. Fold-specific masks and target-edge removal prevent a supervised citation from reappearing as a message-passing feature. Both rankers use the same listwise objective; architecture and optimization settings are in Appendix C.

4.3 Score-and-Rank Fusion and Inference

Score fusion and reciprocal-rank evidence (Cormack et al., 2009) capture different agreement signals. With within-window standardization $z(\cdot)$,

$$s_{\text{score}} = .4z(s_{\text{GAT}}) + .6z(s_{\text{CE}}), \tag{3}$$

$$s_{\text{rank}} = \frac{1}{20 + r_{\text{CE}}} + \frac{1}{20 + r_{\text{GAT}}}, \tag{4}$$

$$s_{\text{final}} = .4z(s_{\text{rank}}) + .6s_{\text{score}}. \tag{5}$$

Candidates are sorted by s_{final}, and the first ten form the recommendation set. The fusion coefficients are selected once during development and then fixed across domains; no domain-specific tuning is performed on the reported test sets.

4.4 Domain-Specific Citation-Sentence Generation

An optional domain-specific Qwen3-32B adapter (Yang et al., 2025) generates one candidate-grounded sentence for each fixed Top-10 reference. Quantized LoRA updates (Hu et al., 2022) are applied to the attention projections using training examples drawn from the fused Top-50: observed citations provide both bundle-level and single-candidate targets, while prompt tokens are

Table 2: End-to-end EWM comparison. Methods use native candidate pipelines; learned retrievers report mean ± standard deviation over three seeds.

Method	Hits@10 ↑	Hits@30 ↑	nDCG@10 ↑	MRR@10 ↑	R@30 ↑	R@100 ↑
BM25	1.1916	2.3018	0.1391	0.3450	0.0802	0.1507
BGE (pretrained)	1.8136	3.3333	0.2111	0.4657	0.1185	0.2180
SciBERT (unfine-tuned)	0.6535	1.3071	0.0795	0.2232	0.0431	0.0863
MasterSet-style SciBERT–NTX	1.8626 ± 0.0219	3.8513 ± 0.0364	0.2056 ± 0.0016	0.4338 ± 0.0025	0.1361 ± 0.0014	0.2595 ± 0.0023
HLM-style GTE retriever	1.1190 ± 0.0452	2.2905 ± 0.0667	0.1280 ± 0.0042	0.3089 ± 0.0126	0.0797 ± 0.0017	0.1549 ± 0.0043
HLM – DeepSeek-V4-Flash-0731	1.5669	–	0.1853	0.4366	–	–
HLM – GLM-5.1	1.6850	–	0.2005	0.4698	–	–
HLM – Qwen3.8-Flash-Next	1.7349	–	0.2099	0.4960	–	–
RW-Cite	**4.0000**	**7.3937**	**0.4445**	**0.6777**	**0.4860**	**0.7730**

Analyzer–Decider rows return only a final Top-10 list and therefore have no deeper-recall values.

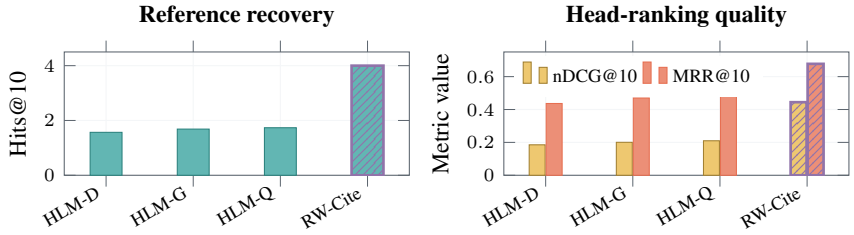

Figure 2: End-to-end HLM comparison. HLM-D/G/Q pair the trained GTE retriever with DeepSeek-V4-Flash-0731, GLM-5.1, and Qwen3.8-Flash-Next, respectively; hatching highlights RW-Cite.

masked from the loss. At inference, title-inconsistent or near-duplicate sentences are regenerated with candidate-specific prompts. The adapter cannot add, remove, or reorder papers, keeping generation separate from ranking evaluation; complete training settings appear in Appendix C.

5 Evaluation

5.1 End-to-End Baseline Comparisons

Setup. We compare complete title–abstract systems under the same corpus, queries, gold labels, temporal controls, metrics, and Top-10 contract. Baselines are BM25 (Robertson & Zaragoza, 2009), pretrained BGE (Xiao et al., 2024), unfine-tuned SciBERT (Beltagy et al., 2019), MasterSet-style SciBERT–NTX (Ratul et al., 2026; Chen et al., 2020), and HLM-style GTE (Hao et al., 2024; Li et al., 2023). Each retains its native retrieval pipeline. Hits@10 is primary; nDCG@10 and MRR@10 measure head ordering, and Recall@30 and Recall@100 diagnose coverage. Learned retrievers report three-seed mean and standard deviation; deterministic and API systems run once. Local experiments used eight 80-GB NVIDIA A800 GPUs, dual Xeon Platinum 8358 CPUs, and 1 TiB RAM (CUDA 12.8; driver 575.57.08); hosted LLM inference used external APIs.

SciBERT–NTX is a shared title–abstract bi-encoder trained with multi-positive NT-Xent. HLM uses a two-stage GTE curriculum, then supplies its temporally admissible Top-30 to Analyzer–Deciders based on DeepSeek-V4-Flash-0731, GLM-5.1, or Qwen3.8-Flash-Next (DeepSeek-AI, 2026; GLM-5-Team, 2026; Qiu et al., 2026). Each LLM receives frozen candidate evidence, cannot retrieve new papers, and must return ten supplied IDs. Full recipes and controls are in Appendix D.

Comparison results. RW-Cite leads every reported metric (Table 2): 4.000 Hits@10, 7.394 Hits@30, 0.445 nDCG@10, 0.678 MRR@10, 0.486 Recall@30, and 0.773 Recall@100. The strongest external deep-retrieval values are 0.1361 ± 0.0014 and 0.2595 ± 0.0023 Recall@30/100 from SciBERT–NTX, showing that the gain begins with candidate discovery. Qwen raises HLM-GTE from 1.119 to 1.735 Hits@10, but remains $2.31\times$ below RW-Cite; Figure 2 confirms the advantage in recovery and head ordering. Deeper recall is unavailable for LLMs that return only ten papers.

Table 3: Multi-domain ranking under the split-masked protocol. Δ is the absolute fusion gain over CE; checkpoints are trained per domain.

Domain	CE Hits@10	GAT Hits@10	Fusion Hits@10	Δ Hits@10	Fusion Hits@30
EWM	3.383	3.480	**4.000**	+0.617	7.394
GW	3.064	3.110	**3.658**	+0.594	7.694
Driving	2.960	2.948	**3.402**	+0.442	6.616
Cosmo	2.865	2.581	**3.256**	+0.391	6.758
Radio	2.411	2.350	**2.975**	+0.564	5.933
WSI	2.454	2.151	**2.822**	+0.368	4.876
Exo	2.492	1.938	**2.769**	+0.277	5.718
SQC	2.345	2.498	**2.727**	+0.382	5.539
FNO	2.260	1.932	**2.630**	+0.370	4.667
SCE	2.088	2.118	**2.569**	+0.481	5.157

Table 4: EWM ablations against Top-400 score-and-rank fusion under the same temporally controlled graph.

Component	Setting	Hits@10 ↑	Δ Hits@10	Hits@30 ↑
Full ranking stack	Score-and-rank fusion, $W = 400$	4.000	–	7.394
Scoring channel	Cross-encoder only	3.383	-0.617	6.598
Scoring channel	GAT only	3.480	-0.520	6.664
Fusion rule	Score fusion without rank fusion	3.924	-0.076	7.268

5.2 Multi-Domain Evaluation

Protocol. We train the same CE, GAT, and fusion pipeline per domain using its masked graph and frozen manifests. CE uses candidate citation contexts and in-window negatives; GAT scores the same structural Top-400. This tests transfer without domain-specific ranking rules; complete metrics are in Appendix A.2.

Results. Fusion leads both individual channels in every domain (Table 3). Its CE gain ranges from $+0.277$ to $+0.617$ (mean $+0.449$), while fused Hits@30 ranges from 4.667 to 7.694. The positive gain in every row shows that the benefit is cross-domain rather than driven by one favorable graph.

5.3 Component Analysis

Protocol. Controlled EWM ablations preserve the masked graph, candidates, and manifest. Against Top-400 score-and-rank fusion, we retain CE or GAT alone or remove the rank-fusion contribution while preserving standardized score fusion. The component rows are evaluated on the same frozen EWM manifest.

Results. Both channels are complementary (Table 4): CE-only and GAT-only reduce Hits@10 by 0.617 and 0.520. Standardized score fusion recovers most of this gap, while adding reciprocal-rank evidence provides the final improvement from 3.924 to 4.000 Hits@10 and from 7.268 to 7.394 Hits@30.

5.4 Recommendation Relevance Beyond Citation Recovery

Protocol. Because bibliographies are incomplete relevance labels, Qwen3-32B judges the frozen Top-30 for 640 training queries as *core*, *related*, *peripheral*, or *none*. It sees query/candidate title–abstract text, but not citation status or scores, and cannot rerank.

Results. Of 19,200 recommendations, 96.2% are core or related (Table 5): 97.8% for gold and 95.1% for non-gold candidates. These judgments suggest that many non-gold recommendations may nevertheless be topically and scientifically relevant. Among gold, 32.1% are core and 65.7% related.

Table 5: LLM relevance assessment of RW-Cite's Top-30. Gold denotes an observed Introduction/Related-Work citation; non-gold is absent from that bibliography.

Scope / coverage	Value	Relevance label	Value	Citation analysis	Value
Queries	640	Core	26.3%	Gold pairs	8,019 (41.8%)
Candidates/query	30	Related	69.9%	Gold core/related	97.8%
Total judgments	19,200	Core or related	**96.2%**	Non-gold core/related	95.1%
Parsed	100.0%	Peripheral	3.6%	$P(\text{core} \mid \text{gold})$	32.1%
Mean gold/query	12.5	None	0.1%	$P(\text{related} \mid \text{gold})$	65.7%
Median gold/query	12	–	–	$P(\text{peripheral} \mid \text{gold})$	2.1%
Zero-gold queries	5	–	–	$P(\text{none} \mid \text{gold})$	$< 0.1\%$

5.5 Downstream Case Study: Citation-Sentence Generation

Protocol. On held-out EWM queries, the adapter generates one sentence per frozen Top-10 reference and may rewrite title-inconsistent or near-duplicate text. We measure validity, identity, grounding, diversity, and citation-context BERTScore (Zhang et al., 2020); ranking is unchanged.

Table 6: Citation-generation quality on held-out EWM queries with fixed reference selection.

Layer	Measurement	Mean ↑	Interpretation
Validity	Nonempty output	1.000	Complete generation
Validity	Candidate-ID consistency	1.000	Output remains attached to its input paper
Validity	Title consistency	0.918	Candidate-specific semantic grounding
Grounding	Title-keyword alignment	0.894	Lexical grounding in the candidate title
Diversity	Pairwise distinctness	0.996	Limited template repetition across papers
Semantic	Citation-context BERTScore F1	0.851	Similarity to observed citation language

Results. Outputs are complete and ID-aligned (Table 6), with 0.918 title consistency, 0.894 keyword alignment, 0.996 distinctness, and 0.851 citation-context BERTScore F1. These results provide an initial demonstration that the fixed recommendations can support candidate-grounded citation-sentence generation; we do not interpret them as evidence of superiority over dedicated generation systems.

6 Limitations

RW-Cite is constructed from arXiv metadata and source documents. Because arXiv coverage varies by field and omits non-deposited work, the ten STEM-centered graphs are not exhaustive, and transfer to other publication and citation cultures remains untested. The supplementary relevance assessment uses LLM judgments without domain-expert review and therefore is not expert-validated annotation. Finally, citation-edge density and local connectivity affect the structural evidence available to shortlisting and graph attention; sparse graphs motivate density-robust retrieval, adaptive fusion, and stronger semantic fallbacks.

7 Conclusion

RW-Cite provides an automated, auditable benchmark for cold-start citation recommendation across ten scientific domains. Its reference framework separates candidate coverage from final ordering and consistently benefits from combining citation-aware textual and graph evidence. On EWM, it reaches 4.000 Hits@10—2.31× the strongest evaluated complete HLM pipeline—while fusion improves over both CE and GAT in every domain. The complementary relevance analysis further identifies useful non-gold prior work beyond the observed bibliographies. By publicly providing the datasets, temporal splits, graphs, contextual evidence, manifests, construction code, and evaluation framework, RW-Cite establishes a reproducible foundation for developing and diagnosing literature-grounding systems.

AI USE STATEMENT

Generative AI tools assisted with literature analysis, experimental planning, software implementation, data analysis, figure preparation, and manuscript editing. The authors reviewed and verified the resulting code, experiments, reported values, citations, figures, and text, and take responsibility for the content of this work.

ETHICS STATEMENT

We have reviewed the ICLR Code of Ethics and designed this work accordingly. RW-Cite uses scholarly metadata and source documents from arXiv and involves no human subjects or private personal data. The benchmark may nevertheless inherit uneven corpus coverage, paper-visibility effects, and field-specific citation practices; observed citations are therefore reproducible supervision rather than exhaustive relevance judgments, and the supplementary LLM labels are not expert-validated. The release documents source provenance and applicable licensing constraints. RW-Cite is intended to assist auditable literature discovery, not to replace expert review or measure scientific merit.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe benchmark construction, the reference framework, evaluation contracts, and the compute environment. Appendices A–E provide data schemas, temporal controls, candidate construction, model and optimizer settings, external-baseline protocols, and metric definitions. Appendix F specifies the released artifacts: versioned manifests bind each result to its metadata snapshot, graph and split checksums, candidate lists, code revision, model configuration, and random seed, while saved per-query rankings permit independent metric recomputation. We publicly provide the benchmark data and the construction, training, and evaluation code.

REFERENCES

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 238–251, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1022. URL https://aclanthology.org/N18-1022/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759. Association for Computing Machinery, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. doi: 10.48550/arXiv.2606.19348. URL https://arxiv.org/abs/2606.19348.

Michael Färber and Adam Jatowt. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405, 2020. doi: 10.1007/s00799-020-00288-2. URL https://doi.org/10.1007/s00799-020-00288-2.

GLM-5-Team. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. doi: 10.48550/arXiv.2602.15763. URL https://arxiv.org/abs/2602.15763.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Qianyue Hao, Jingyang Fan, Fengli Xu, Jian Yuan, and Yong Li. HLM-Cite: Hybrid language model workflow for text-based scientific citation prediction. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. doi: 10.52202/079017-1527. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5635925cf9d2274f338eb0dd5971e845-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiangci Li, Biswadip Mandal, and Jessica Ouyang. CORWA: A citation-oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5426–5440, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.397. URL https://aclanthology.org/2022.naacl-main.397/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. doi: 10.48550/arXiv.2308.03281. URL https://arxiv.org/abs/2308.03281.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://www.nature.com/articles/s41586-026-10265-5.

Michael H. MacRoberts and Barbara R. MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342–349, 1989.

Sheshera Mysore, Arman Cohan, and Tom Hope. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4453–4470, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.331. URL https://aclanthology.org/2022.naacl-main.331/.

Jeppe Nicolaisen. Citation analysis. *Annual Review of Information Science and Technology*, 41(1): 609–641, 2007. doi: 10.1002/aris.2007.1440410120.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. doi: 10.48550/arXiv.1901.04085. URL https://arxiv.org/abs/1901.04085.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11670–11688, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.802. URL https://aclanthology.org/2022.emnlp-main.802/.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026. doi: 10.48550/arXiv.2608.30320. URL https://arxiv.org/abs/2608.30320.

Md Toyaha Rahman Ratul, Zhiqian Chen, Kaiqun Fu, Taoran Ji, and Lei Zhang. MasterSet: A large-scale benchmark for must-cite citation recommendation in the AI/ML literature. In *Proceedings of the 2026 SIAM International Conference on Data Mining (SDM)*, 2026. URL https://arxiv.org/abs/2604.17680. Forthcoming.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Andreas Varvarigos, Ali Maatouk, Jiasheng Zhang, Ngoc Bui, Jialin Chen, Leandros Tassiulas, and Rex Ying. LitBench: A graph-centric large language model benchmarking tool for literature tasks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 1*, pp. 2806–2817, Jeju Island, Republic of Korea, 2026. Association for Computing Machinery. doi: 10.1145/3770854.3785679. URL https://doi.org/10.1145/3770854.3785679.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, Washington, DC, USA, 2024. Association for Computing Machinery. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

APPENDICES

A BENCHMARK CONSTRUCTION AND DATA CONTRACTS

This appendix specifies the data, temporal, modeling, and evaluation contracts used for the reported multi-domain and external-baseline results.

Domain retrieval configuration. Each domain is defined by a versioned set of prototype phrases covering its principal research themes. Prototype selection and retrieval parameters are fixed before citation extraction and do not use test labels or ranking results. The released manifest records the prototype text, encoder checkpoint, corpus snapshot, similarity threshold, rank cap, and retained seed identifiers, allowing every domain corpus to be reconstructed independently.

Table 7: Complete fused-framework results across the ten domain benchmarks.

Domain	Hits@10	Hits@30	nDCG@10	MRR@10	R@30	R@100
EWM	4.000	7.394	0.445	0.678	0.486	0.773
GW	3.658	7.694	0.406	0.652	0.404	0.743
Driving	3.402	6.616	0.398	0.663	0.460	0.771
Cosmo	3.256	6.758	0.380	0.649	0.422	0.702
Radio	2.975	5.933	0.357	0.592	0.412	0.725
WSI	2.822	4.876	0.375	0.653	0.520	0.783
Exo	2.769	5.718	0.338	0.570	0.459	0.734
SQC	2.727	5.539	0.331	0.599	0.428	0.736
FNO	2.630	4.667	0.353	0.628	0.492	0.744
SCE	2.569	5.157	0.325	0.571	0.461	0.740

A.1 Domain-Graph Schema and Gold Labels

For each domain d, the benchmark stores a directed graph $\mathcal{G}_d = (V_d, E_d)$. A node represents one resolved paper and must contain a canonical identifier, title, abstract when available, publication date, and source metadata. An edge (u, v) records that paper u cites paper v in its Introduction or a Related Work-like section. When extraction succeeds, the edge also stores the local citation sentence $\xi_{u,v}$, its source section, and the raw and cleaned forms of the text.

For a query q, the frozen full-gold set is

$$G_q^{\text{full}} = \{c : (q, c) \in E_d,\ c \text{ is resolved}\}. \tag{6}$$

All methods within a result table are evaluated against the same frozen G_q^{full} labels. The benchmark does not truncate G_q^{full} to ten citations. Consequently, Hits@10 is a count of recovered gold papers, while $\text{R@10}_{\text{full}}$ divides by the complete frozen gold set. These labels represent citations observed in the query's Introduction and Related Work sections. They are not exhaustive human judgements of every paper that could be useful to cite.

A.2 Complete Multi-Domain Metrics

Table 7 complements the main-text component comparison with head-ordering and recall metrics for the fused framework. The metrics are computed independently on each domain's frozen test split.

A.3 Temporal Eligibility and Leakage Audit

RW-Cite orders admitted queries by publication month, assigns the newest approximately ten percent to test, and reserves a one-month corpus-only frontier. The scoring graph is produced by removing outgoing edges from test and frontier sources. Test-query outgoing citation sentences and edges are therefore labels only and are not inference features. Accordingly, our protocol should be interpreted as split-level leakage control rather than a strict per-query historical graph reconstruction.

The released benchmark split contains disjoint training and test queries together with a corpus-only frontier. Development subsets, when used for model or configuration selection, are drawn exclusively from the training period and remain disjoint from the test queries. The structural shortlister reserves a random 15% of the training queries as an internal development subset using seed 0. The external comparison instead uses a chronological, whole-month development partition drawn before the test period; records with unknown dates remain training-only. Test queries are excluded from structural-model, cross-encoder, and GAT training and from hyperparameter selection. All known gold identifiers are removed from the negative set for their query, including positives not sampled in the current optimization step. Query self-links and duplicate candidate identifiers are also removed.

The benchmark artifact bundle records query lists, graph fingerprints, model configurations, and per-query predictions.

Every run must save an audit manifest containing the following fields:

- train, optional development, and test query-ID hashes;
- graph, split, source-list, and candidate-manifest checksums;
- query and candidate publication dates and the applied graph mask;
- IDs removed for self-match, split ineligibility, duplicate identity, or full-gold membership;
- code revision, model checkpoint, configuration, and random seed.

A.4 Citation-Context Cleaning and Serialization

Citation-context cleaning removes comments, labels, layout commands, figure and table fragments, malformed prefixes, and excessive citation key lists without changing graph topology.

The query is serialized as its title followed by its abstract. An RW-Cite candidate is serialized as

$$x_c = [t_c; \xi_1^{c,\mathrm{clean}}; \ldots; \xi_m^{c,\mathrm{clean}}], \qquad m \le 3. \tag{7}$$

If no incoming context is available in the scoring graph, the representation falls back to the candidate title and a short abstract fragment. The exported text-view manifest records the selected edge identifiers, serialized text, token count, truncation status, and fallback status.

B Candidate-Universe and Structural-Support Details

B.1 Candidate-Universe Union

Let D_q be the deduplicated union of multi-view BGE retrieval and pseudo-relevance-feedback results, H_q^{hub} the query-specific topic hubs, and $P_q^{(2)}$ the candidates selected for weak second-hop propagation. The candidate universe is

$$U_q = \mathrm{Dedup}\left(D_q \cup \bigcup_{h \in H_q^{\mathrm{hub}}} \mathrm{Out}(h) \cup \bigcup_{p \in P_q^{(2)}} \mathrm{Out}(p) \right) \setminus \{q\}. \tag{8}$$

Candidate construction uses the split-masked graph described in Section 3.5. Gold labels do not participate in the construction or ordering of U_q.

B.2 Hub and Propagated Support

Let T_q contain the first 60 papers in the ordered dense-retrieval results. We construct topic hubs from the first 120 dense results with out-degree at least two and from papers that cite those results, also subject to a minimum out-degree of two. After deduplication, the hub set is capped at 400 papers.

For each topic hub h, let

$$k_q(h) = |\mathrm{Out}(h) \cap T_q| \tag{9}$$

and assign the weight

$$\omega_q(h) = \frac{g(k_q(h))}{\log_2(2 + d_{\mathrm{out}}(h))}, \tag{10}$$

where

$$g(k) = \begin{cases} 0.12, & k = 0, \\ 1.5, & k = 1, \\ 2.5, & k \ge 2. \end{cases} \tag{11}$$

The first-hop support assigned to candidate c is

$$S_q^{(1)}(c) = \sum_{h \in H_q} \mathbb{I}[(h, c) \in E]\, \omega_q(h). \tag{12}$$

Let P_q contain the 100 highest-support candidates satisfying $S_q^{(1)}(p) \ge 0.5$ and $d_{\mathrm{out}}(p) \ge 1$. The weaker second-hop support is

$$S_q^{(2)}(c) = \sum_{p \in P_q} \mathbb{I}[(p, c) \in E] \frac{0.2}{\log_2(2 + d_{\mathrm{out}}(p))}. \tag{13}$$

The total support is $S_q(c) = S_q^{(1)}(c) + S_q^{(2)}(c)$. The focused-hub set is

$$F_q = \{h \in H_q : |\operatorname{Out}(h) \cap T_q| \geq 1\}, \tag{14}$$

and $C_q(c)$ is the normalized number of focused hubs that cite c.

B.3 Structural Shortlisting

For each $c \in U_q$, the structural ranker forms the ten-dimensional feature vector

$$\begin{aligned}\boldsymbol{\phi}(q,c) = [&S_q(c), \log(1+S_q(c)), J(q,c), C_q(c),\\ &\log(1+d_{\text{in}}(c)), \log(1+d_{\text{out}}(c)), I_D(c), B_q(c), I_S(c), e_q^\top e_c].\end{aligned} \tag{15}$$

where I_D indicates dense-retrieval membership, B_q is a dense-rank bonus, I_S indicates nonzero graph support, and $e_q^\top e_c$ is embedding similarity. Features are standardized using training statistics and a logistic model estimates $p_{\text{struct}}(q,c) = \sigma(b + \mathbf{w}^\top \boldsymbol{\phi}(q,c))$. The 400 highest-scoring candidates form W_q. The benchmark stores the universe and ordered-window manifests together with their coverage statistics.

C RW-Cite Model and Training Details

C.1 SciBERT Cross-Encoder

The ranker is initialized from `allenai/scibert_scivocab_uncased` with a newly initialized scalar sequence-classification head. Each query–candidate pair is jointly tokenized and truncated to 256 tokens. A listwise step contains up to 64 candidates from one query, including all sampled positives up to one quarter of the list and negatives filling the remaining slots. For a sampled list L_q with included positive set P_q', the listwise objective is

$$\mathcal{L}_q = -\frac{1}{|P_q'|} \sum_{c \in P_q'} \log \frac{\exp s_{\text{CE}}(q,c)}{\sum_{c' \in L_q} \exp s_{\text{CE}}(q,c')}. \tag{16}$$

Table 8 records the complete cross-encoder training and inference configuration.

Table 8: RW-Cite cross-encoder configuration.

Setting	Value
Backbone	SciBERT, fresh scalar head
Reranking window / list size	400 / 64
Maximum sequence length	256
Maximum incoming contexts	3
Optimizer / weight decay	AdamW / 0.01
Body learning rate	2×10^{-5}
Warmup ratio	0.06
Training	2 epochs; random in-window negatives
Inference window	learned structural Top-400
Cross-encoder output	$s_{\text{CE}}(q,c)$

C.2 Graph-Attention Model and Fusion

The graph-attention model uses SciBERT node embeddings and sparse incoming-citation, outgoing-citation, and co-citation relations. Two graph-attention layers produce $s_{\text{GAT}}(q,c)$ over the same Top-400 window. Fold masks and target-edge removal are applied during training. At inference, the specified score-and-rank fusion configuration combines cross-encoder and GAT outputs. For scores standardized within a candidate window, score fusion is

$$s_{\text{score},\alpha}(q,c) = \alpha z(s_{\text{GAT}}(q,c)) + (1-\alpha) z(s_{\text{CE}}(q,c)). \tag{17}$$

Let $r_{\mathrm{CE}}(q, c)$ and $r_{\mathrm{GAT}}(q, c)$ be the corresponding ranks. Rank fusion is

$$s_{\mathrm{rank},k}(q, c) = \frac{1}{k + r_{\mathrm{CE}}(q, c)} + \frac{1}{k + r_{\mathrm{GAT}}(q, c)}, \tag{18}$$

and the final score is

$$s_{\mathrm{final}}(q, c) = wz(s_{\mathrm{rank},k}(q, c)) + (1 - w)s_{\mathrm{score},\alpha}(q, c). \tag{19}$$

We use $\alpha = 0.4$, rank-fusion offset $k = 20$, and fusion weight $w = 0.4$.

C.3 Citation-Generation Adapter

Table 9 records the adapter configuration. For every training query, the fused Top-50 pool is intersected with observed training citations. Each valid source contributes two bundle examples and two single-candidate examples. Prompts are masked from the language-model loss, so optimization is applied only to the assistant response. At evaluation, all reported fractions are computed per query and then macro-averaged. Pairwise distinctness is the fraction of unordered sentence pairs that do not cross the near-duplicate threshold of 0.8; BERTScore is evaluated only where an observed citation-context sentence is available.

Table 9: Citation-generation adapter configuration.

Setting	Value
Base model	Qwen3-32B
Parameter-efficient update	LoRA on q/k/v/o projections
LoRA rank / scaling / dropout	16 / 64 / 0.05
Quantization / compute precision	8-bit / bfloat16
Maximum sequence length	2048
Optimizer schedule	10^{-4} cosine; 50 warmup steps
Epochs / batch / accumulation	2 / 1 / 8
Training candidate pool	Fused Top-50
Inference input	Fixed ranked Top-10
Hidden reasoning	Disabled

D External Comparison Baselines

D.1 SciBERT–NTX

SciBERT–NTX is our adapted MasterSet-style dense baseline (Ratul et al., 2026). A shared SciBERT encoder independently maps the query and candidate title–abstract to attention-mask mean-pooled, ℓ_2-normalized vectors. Their similarity is

$$s_{\mathrm{NTX}}(q, c) = z_q^\top z_c. \tag{20}$$

For a minibatch of queries and sampled positives, the multi-positive loss is

$$\mathcal{L}_q^{\mathrm{NTX}} = -\frac{1}{|P'_q|} \sum_{p \in P'_q} \log \frac{\exp(s_{\mathrm{NTX}}(q, p)/\tau)}{\sum_{c \in C'_q} \exp(s_{\mathrm{NTX}}(q, c)/\tau)}. \tag{21}$$

If a sampled candidate is also a gold citation for another query in the same batch, it is included in that query's positive mask rather than treated as a false negative. The trained encoder performs exact cosine search over every temporally legal EWM candidate and returns Top-10. Table 10 summarizes the configuration used for every SciBERT–NTX seed.

We additionally evaluate the same retrieval procedure with the unfine-tuned SciBERT checkpoint. This control separates the contribution of scientific pretraining from EWM-specific contrastive optimization.

Table 10: Configuration of the adapted MasterSet-style baseline.

Setting	Value
Encoder	`allenai/scibert_scivocab_uncased`
Input	title + abstract
Pooling / similarity	masked mean / cosine
Maximum length	256 tokens per paper
Temperature τ	0.05
Optimizer / weight decay	AdamW / 0.01
Learning rate / warmup	2×10^{-5} / 0.10
Training epochs	5
Physical batch / accumulation	32 / 4
Effective optimization batch	128
Seeds	42, 43, 44

D.2 HLM-Style Retriever

The HLM-style baseline adapts the curriculum-trained retrieval component of HLM-Cite (Hao et al., 2024). Core, superficial, and non-citation labels are constructed from the training data. Stage 1 uses core citations and in-batch negatives; Stage 2 uses graded relevance $(2, 1, 0)$ and a smooth NeuralNDCG objective. Both stages initialize GTE-base, freeze the bottom seven layers, use a maximum length of 512 tokens, and train for ten epochs.

We first evaluate the resulting GTE retriever directly over the complete eligible corpus. We then evaluate complete HLM-based pipelines in which the trained retriever supplies its temporally admissible Top-30 candidates and each hosted Analyzer–Decider selects the final ten references. The LLM receives the query title and abstract, frozen candidate evidence, and the fixed training-only one-shot demonstration. It cannot retrieve additional papers and must select exactly ten unique identifiers from the supplied window. Results are reported in Table 2 and Figure 2.

D.3 Unified End-to-End Candidate Protocol

The external comparison evaluates complete systems under one end-to-end contract. BM25, pretrained encoders, SciBERT–NTX, and the HLM-style GTE model search the eligible EWM corpus using their native retrieval scores. The complete HLM pipelines pass the trained GTE retriever's temporally admissible Top-30 candidates to their respective Analyzer–Decider models, while RWCite uses its native universe construction, structural Top-400, neural scoring, and fusion. Candidate sets may therefore differ because candidate generation is part of each framework. The query list, eligible corpus, full-gold labels, evaluation implementation, and Top-10 output contract are held fixed, making Hits@10 directly comparable across systems. Exact candidate manifests and per-query outputs are included in the artifacts that we publicly provide.

E Evaluation Metrics and Statistical Analysis

Let $R_q@k$ be the first k items in a model ranking. The primary metric is the average number of gold citations recovered in the first ten:

$$\text{Hits@10} = \frac{1}{|Q|} \sum_{q \in Q} |G_q^{\text{full}} \cap R_q@10|. \tag{22}$$

This count must not be described as a binary hit rate. Full-gold recall is

$$\text{R@10}_{\text{full}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|G_q^{\text{full}} \cap R_q@10|}{|G_q^{\text{full}}|}. \tag{23}$$

With binary relevance, we additionally report

$$\mathrm{DCG@10}_q = \sum_{i=1}^{10} \frac{\mathbb{I}[r_{q,i} \in G_q^{\mathrm{full}}]}{\log_2(i+1)}, \tag{24}$$

$$\mathrm{nDCG@10} = \frac{1}{|Q|} \sum_q \frac{\mathrm{DCG@10}_q}{\mathrm{IDCG@10}_q}, \tag{25}$$

$$\mathrm{MRR@10} = \frac{1}{|Q|} \sum_q \frac{\mathbb{I}[i_q \le 10]}{i_q}, \tag{26}$$

where $i_q = \min\{i : r_{q,i} \in G_q^{\mathrm{full}}\}$ and a query with no gold in the first ten contributes zero.

For a method-specific candidate pool C_q, we report its full-gold coverage

$$\mathrm{Coverage}(C) = \frac{1}{|Q|} \sum_q \frac{|G_q^{\mathrm{full}} \cap C_q|}{|G_q^{\mathrm{full}}|} \tag{27}$$

This quantity separates candidate-generation failures from ordering errors.

Training-based external retrievers are run with seeds 42, 43, and 44 and summarized by mean and standard deviation. This includes SciBERT–NTX and the trained retrieval component of the HLM-style pipeline. Per-query predictions are retained for every run. Each Analyzer–Decider uses a fixed model identifier, temperature of zero, and the same training-only one-shot demonstration.

F Reproducibility Checklist and Artifacts

We publicly provide the benchmark datasets, temporal splits, domain graphs, candidate manifests, contextual evidence, trained baseline configurations, per-query outputs, and the code required for construction, training, and evaluation. Versioned manifests associate each reported result with its metadata snapshot, query split, graph checksum, model configuration, and random seed.

Benchmark versioning. Each benchmark release is immutable and identified by its metadata snapshot, graph checksum, split manifest, and candidate-manifest version. A corpus refresh creates a new version while preserving prior releases and their evaluation artifacts. Results from different versions are reported separately unless they are recomputed on an identical query and candidate manifest.

The artifact bundle associated with each result table contains:

- frozen train, optional development, and test query lists and graph checksums;
- per-query full-gold IDs and recorded temporal eligibility decisions;
- the legal corpus, dense-retrieval outputs, and RW-Cite universe/Top-400 manifests with coverage statistics;
- cleaned citation-context text with provenance and truncation metadata;
- citation-generation adapter weights, fixed input selections, generated sentences, rewriting decisions, and per-query quality metrics;
- model checkpoint identifiers, complete hyperparameters, training logs, seeds, and optimizer-update counts;
- SciBERT–NTX embeddings or exact regeneration configuration;
- per-query complete rankings, scores when available, gold flags, and aggregate metrics;
- hardware, software versions, code revision, runtime, memory, and monetary cost records.

For every comparison table, the evaluation script verifies that all methods within the same setting contain exactly the same query IDs, that a fixed-window comparison contains identical candidate IDs, that no test query appears in a training artifact, and that every metric can be recomputed from the saved per-query rankings. Temporal eligibility is checked under the protocol declared for that table and recorded in the corresponding manifest.

Table 11: Construction and use of the fixed Analyzer–Decider demonstration.

Property	Specification
Source partition	Training only
Query input	Title and abstract
Candidate input	Frozen training Top-30
Candidate evidence	Identifier, title, and up to three citation sentences
Analyzer target	JSON assessment for all 30 candidate identifiers
Decider target	JSON list containing exactly ten unique identifiers
Target supervision	Observed citations of the training query
Human intervention	One-time review before freezing
Reuse across models	Identical demonstration for every evaluated LLM
Excluded information	Test data, retrieval scores, and explicit gold indicators

Table 12: Inference configuration for the HLM-based comparison.

Setting	Analyzer	Decider
Temperature	0	0
Maximum output tokens	2,048	512
Required coverage	All 30 candidates	Exactly ten candidates
Output format	JSON object	JSON ranked-ID list
External retrieval	Disabled	Disabled
Maximum retries	3	3
Hidden reasoning	Disabled when supported	Disabled when supported

G Analyzer–Decider Prompting and Downstream Integration

G.1 Fixed One-Shot Demonstration

All HLM-based Analyzer–Decider comparisons use one fixed, manually reviewed demonstration drawn exclusively from the training partition. The example contains one query title and abstract and a frozen set of 30 training candidates represented in the same form used at evaluation: candidate identifier, title, and up to three temporally admissible incoming citation sentences, with the abstract used when citation evidence is unavailable.

The Analyzer portion maps every candidate identifier to a concise assessment of its possible citation role. The Decider portion contains exactly ten unique candidate identifiers. These targets are selected from observed citations of the training query, but gold membership is not displayed as an input feature. The demonstration excludes development and test queries, retrieval scores, test labels, and information absent from the supplied evidence. After manual review, its serialized content and checksum are frozen and reused unchanged for every evaluated LLM. Table 11 summarizes the construction and information controls applied to this demonstration.

The Analyzer is instructed to use only the supplied query and candidate evidence, to assess every candidate without ranking or omission, and to return a JSON mapping from candidate identifiers to analyses. The Decider receives the same query and candidate evidence together with the Analyzer output. It must select exactly ten unique identifiers from the supplied candidates and return only `{"ranked_ids":["id1",...,"id10"]}`. The decision criteria prioritize directly reused methods, experimental baselines, datasets, evaluation protocols, foundational work, and closely related solutions. Table 12 records the common decoding and output constraints used by the Analyzer and Decider.

Table 13: Use of RW-Cite outputs within a downstream research agent.

Stage	RW-Cite output	Agent use
Query routing	Domain-specific configuration	Selects the appropriate literature graph and models
Retrieval	Query-specific candidate universe	Establishes the auditable search space
Shortlisting	Ordered structural Top-400	Defines the shared neural-ranking window
Ranking	Fused Top-10 references	Supplies related-work recommendations
Evidence	Candidate metadata and citation contexts	Supports inspection and citation grounding
Generation	Candidate-aligned citation sentences	Produces related-work prose for fixed references
Provenance	Graph, split, candidate, and model identifiers	Enables reconstruction and auditing

G.2 Downstream Research-Agent Integration

A downstream research-agent system can use RW-Cite as its literature-discovery and citation-grounding component. Given a proposed study's title and abstract, the agent selects the appropriate scientific domain and submits the query to the corresponding RW-Cite pipeline. RW-Cite constructs the candidate universe, applies learned structural shortlisting, scores the shared Top-400 window with the cross-encoder and graph-attention components, and returns the ten highest-ranked references under score-and-rank fusion.

The integration preserves the boundary between recommendation and generation. The agent consumes the ordered Top-10 identifiers, paper metadata, candidate-side evidence, and provenance records without changing their order. When citation prose is requested, the domain-specific citation-generation adapter receives these fixed references and generates one candidate-grounded sentence per paper. The adapter cannot introduce a new reference, remove a selected reference, or alter the ranking. Table 13 summarizes how each RW-Cite output is consumed without altering the evaluated ranking.

The downstream agent stores the query, selected domain, ranked paper identifiers, candidate-evidence provenance, and model configuration associated with each request. This record allows a recommendation or generated citation sentence to be traced back to the exact graph, candidate window, and model outputs that produced it. This integration is a downstream application of RW-Cite rather than an additional benchmark condition; all ranking results remain those obtained under the frozen evaluation protocol defined in the main paper.