MUSES: RETRIEVING THE LITERATURE THAT SHAPES A RESEARCHER'S NEXT PAPER

Anonymous authors
Paper under double-blind review

ABSTRACT

Much of the work that shapes a researcher's next paper is not something they went looking for: it often arrives before the project has a question anyone could search for. How much of it can be anticipated from a researcher's past publications alone? We introduce MUSES, a benchmark of 1.04 million cases built from the Semantic Scholar Open Research Corpus (S2ORC), of which 140,278 form the held-out test split. Given a researcher's papers before a cutoff, a system ranks 2.33 million earlier candidates for their next paper and is scored primarily on works the researcher had not cited before. Because a citation is weak evidence of influence, we also asked authors directly which works shaped their papers, yielding 580 author-collected evaluable cases, a distinct construct from the citation-based targets. A classifier grounded in human coding adds citation contexts where the next paper builds on a work. Across retrievers from popularity to dense encoders and a supervised model, we find that a researcher's past reveals where their next paper will look, but least well what will be new to them. History-based retrievers place a newly cited work in their top 100 for 29–34% of papers, against 5% for popularity; among author-collected cases, one retriever finds a work the authors themselves named in 56% of cases (36% when that work was new to the researcher). Yet the signal fades for works new to the researcher, most of all for those that shaped the paper: for 46% of the 6,707 test papers with such labels, no retriever we evaluate finds one in its top 1,000. Much of the remaining gap is the project itself: adding only the next paper's title raises one retriever's reach on these papers from 29% to 48%, more than switching retrievers does. MUSES shows where a researcher's record runs out and the project itself must take over.

1 INTRODUCTION

Much of the work that shapes a researcher's next paper is not something they went looking for. It arrives through a conversation, a talk, or a paper stumbled on while looking for something else, before the project has a question anyone could search for. Tools that recommend literature try to make such encounters less a matter of luck. Early in a project, whether they are personalized feeds (Flicke et al., 2025) or ideation agents that retrieve prior papers as inspiration (Wang et al., 2024; Si et al., 2025), they can always draw on one input: the researcher's own publication record. How much that record reveals about the literature of the researcher's next paper is an open question, and its answer bounds what such tools can offer.

Several lines of work address parts of this problem, but each replaces it with an easier proxy. *Researcher-history recommenders* profile a scholar's interests from past papers or ratings (Sugiyama & Kan, 2010; Jiang et al., 2023; Flicke et al., 2025). They are scored on relevance to those interests or on recently cited papers, so a system that resurfaces familiar work scores well. *Citation recommenders* and *literature search* need a draft, a citation context, or an explicit question (He et al., 2010; Ebesu & Fang, 2017; Zheng et al., 2026; Zhang & Ma, 2026; Ajith et al., 2024), all of which exist only after the project has taken shape. *Ideation systems* retrieve prior work as inspiration but are judged on the ideas they generate (Wang et al., 2024; Radensky et al., 2024; Si et al., 2025), not on whether a retrieved work is one a researcher actually goes on to use; ResearchBench checks retrieved inspirations against real papers, but only after giving the system each paper's research question (Liu et al., 2026). Prompting a large language model with the researcher's record is a natural alternative, but it cannot be made time-safe: a generative model may have memorized the very paper it is asked

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

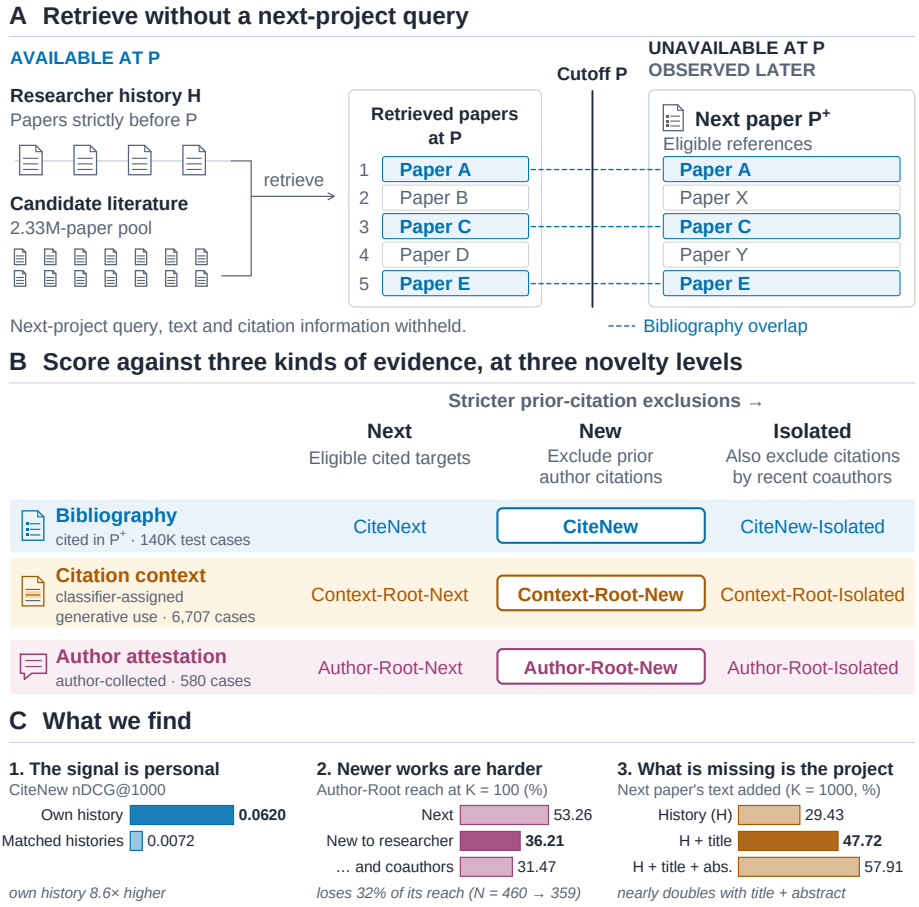

Figure 1: **MUSES at a glance: task, scoring, and findings.** (A) At focal paper P, a retriever sees only the researcher's history H and candidates published before P. (B) Rankings are scored against three kinds of evidence from the next paper P^+, each at three novelty levels; boxed cells are the primary targets. (C) The three findings (§4.1–4.3); the third is a hindsight test that adds P^+'s title and abstract to H.

to anticipate and can reproduce its bibliography, whereas an embedding model only scores the similarity of the texts it is given. Each approach therefore either depends on a project that already exists or is never tested against what a researcher's next paper actually builds on; meeting that bar means testing retrieval against the next paper itself, with everything about it withheld (Table 1).

Meeting that bar is harder than modeling current interests, because research does not always continue in a straight line, and it requires closing three shortcuts that the work above leaves open: seeing the future, resurfacing work the researcher already knew, and counting every citation as influence. We introduce **MUSES**, a benchmark for *pre-project retrieval* named for the Muses, the sources of inspiration it tries to anticipate. MUSES closes each shortcut: a strict time cutoff removes anything published after the project, targets exclude previously cited works, and evidence of influence goes beyond citation.

In our benchmark, each case fixes a researcher, a focal paper P, and their next paper P^+. A system sees only the researcher's publications before P and ranks candidates published before P; the next project's title, text, collaborators, and references stay hidden (Figure 1A). Built from S2ORC (Lo et al., 2020), MUSES contains about 1.04 million cases across fields, with author-disjoint temporal splits (705,476 training, 193,026 validation, and 140,278 test cases) and a fixed pool of 2.33 million candidates. Targets are the works P^+ cites, at three levels of novelty: all of them; those the researcher had not cited before, the primary target; and those no recent coauthor had cited either, which rules out works a collaborator could have passed along.

Table 1: **Only MUSES evaluates retrieval before a project exists, excludes works the researcher already knew, and scores against direct evidence of influence.** Rows group citation recommendation and search, research ideation, and researcher-history recommendation. The last three columns are the requirements of §1; (✓) marks weaker related evidence, namely researchers' relevance judgments or inspirations extracted from a paper. Properties as each source specifies: SciDocs (Cohan et al., 2020), RELISH (Brown et al., 2019), LitSearch (Ajith et al., 2024), CiteRAG (Zheng et al., 2026), ResearchBench (Liu et al., 2026), Sugiyama & Kan (2010), and TAPRec (Jiang et al., 2023).

Benchmark	System input	Candidates	Evaluated against	Pre-project input	Excludes known	Influence evidence
SciDocs (cite)	the citing paper	30 per query	its own references	×	×	×
RELISH	a seed article	ranked lists	relevance to seed	×	×	(✓)
LitSearch	literature question	64K papers	matching papers	×	×	×
CiteRAG	abstract or draft	554K papers	its own references	×	×	×
ResearchBench	question + survey	75 per target	its inspirations	×	×	(✓)
Sugiyama & Kan	own papers	597 papers	relevance to interests	✓	×	(✓)
TAPRec	dated own papers	sampled	recently cited papers	✓	×	×
MUSES	own papers only	2.33M papers	next paper's new references	✓	✓	✓

For the question of which works shaped a paper, there is no substitute for asking its authors, evidence that is rare in retrieval benchmarks because it means reaching working scientists directly, not scraping public text. We contacted the corresponding authors of papers across 25 journals and adjudicated 1,482 replies into 580 evaluable cases of reported influence, a distinct, author-collected set. MUSES pairs this evidence with a scalable complement: citation contexts in P^+ that describe building on a cited work's idea, framing, or method, labeled by a classifier grounded in 1,202 human-coded contexts; 6,707 test cases have such a target in the candidate pool. The two are complementary rather than interchangeable: authors report what shaped their paper, whereas citation contexts show how the paper uses a work (Figure 1B).

We evaluate retrievers from popularity to dense encoders and a supervised model. The results show that **a researcher's past reveals where their next paper will look, but least well what will be new to them**. History-based retrievers place a newly cited work in their top 100 for 29–34% of cases, against 5% for popularity, and matched histories of other researchers retain only about a tenth of the signal (nDCG@1000). Against the most direct evidence, one retriever finds a work the authors themselves named in 56% of the 580 author-attested cases (95% CI 52–60%). The signal fades as works become newer to the researcher, most of all for works that shaped the paper. For 46% of next papers that build on new works in the candidate pool, no retriever we evaluate places one in its top 1,000. **Newness, not obscurity, is what makes works hard to find**: works no one had cited before are found more often than cited ones. For representation learning, several history centroids beat one vector, a general-purpose embedder beats citation-trained ones, and supervision on MUSES does not help: the limit is information more than representation.

What history alone cannot supply is the project itself. In a hindsight test, adding only the next paper's title to the same history raises the share of those papers reached in the top 1,000 from 29% to 48%, more than the 12-point spread across history-based retrievers. Tools that ask researchers for a short description of their next project, such as a working title, could therefore recover a substantial share of what history alone cannot. **MUSES measures where that boundary lies** (Figure 1C).

2 Related Work

Citation recommendation and literature search. Context-aware recommenders use the words around a citation (He et al., 2010; Ebesu & Fang, 2017), CiteRAG uses target-paper and manuscript context (Zheng et al., 2026), CBERT4REC an incomplete manuscript (Zhang & Ma, 2026), and LitSearch an explicit literature question (Ajith et al., 2024). Each needs an input that does not exist before a project does; Table 1 compares them with MUSES.

Recommending from a researcher's past work. Sugiyama & Kan (2010) derive interests from a researcher's recent work, TAPRec models changing interests from historical publications, with coauthor signals for cold-start users (Jiang et al., 2023), and Scholar Inbox personalizes a continuing literature feed from user ratings (Flicke et al., 2025); each is judged on relevance to a history or on recently cited papers, not on what a specific later paper uses. Scientific-document encoders learn from citation structure (Cohan et al., 2020; Ostendorff et al., 2022; Singh et al., 2023), whereas general-purpose embedders are instruction-tuned for broad text similarity (Muennighoff et al., 2025; Zhang et al., 2025); MUSES evaluates both encoder families, and one-vector versus several-interest history summaries, against that outcome (Table 1).

Literature-based discovery and research ideation. Literature-based discovery treats connections latent across literatures as undiscovered knowledge (Swanson, 1986; Smalheiser, 2012). Ideation work also builds knowledge bases of recombined ideas (Sternlicht & Hope, 2026). ResearchBench evaluates inspiration retrieval as part of hypothesis generation from an explicit research question and background (Liu et al., 2026), SciPaths forecasts enabling prior-work pathways given a target scientific contribution (Chamoun et al., 2026), and benchmarks such as DiscoveryBench and AstaBench likewise evaluate discovery and research capabilities once a task or goal has been specified (Majumder et al., 2025; Bragg et al., 2026); each of these is given some description of the project it serves. MUSES evaluates retrieval before one exists.

Evidence of what a citation does. Citation-intent, citation-frame, and citation-function research classifies the roles citations play in published text (Cohan et al., 2019; Jurgens et al., 2018; Arnaout et al., 2026; Vodička et al., 2026), and Valenzuela et al. (2015) identify citations judged meaningful to the citing paper from their context. These classify a citation once written; MUSES instead makes generative use a retrieval target, to be found before any citation exists. Author-report research asks a related question: Wakeling et al. (2025) ask authors how their own work was cited, a matter of practice, whereas MUSES asks which prior works shaped the paper, a matter of influence. The two evidence sources in MUSES therefore remain separate rather than interchangeable ground truth.

3 THE MUSES BENCHMARK

MUSES asks a retriever to anticipate the literature of a researcher's next paper using only what the researcher has already published. This section defines the task, what counts as a correct answer, the evidence that a work shaped a paper, how cases are built, and the retrievers, metrics, and diagnostics.

3.1 TASK: RANKING LITERATURE FROM A RESEARCHER'S PAST WORK

The task models the moment before a project has a question: a system knows the researcher's publication record and nothing about the next paper. Each *case* pairs a researcher's *focal paper* P with the next paper they lead, P^+, both drawn from S2ORC (Lo et al., 2020). We follow researchers across papers on which they are last author, so consecutive papers approximate one continuing research program. The history H holds the researcher's earlier papers in any author position; P sets the cutoff.

A retriever maps the history to a ranking of candidates, $f(H(a,t),\, C(t))$, where a is the researcher, t is the date of P, and the candidate set $C(t) = \{d \in \mathcal{G} : \mathrm{pubdate}(d) < t\}$ contains the papers of a fixed pool $\mathcal{G}$ (§3.4) published at least one day before P. Everything about the next project stays hidden: its title, abstract, text, references, and citation contexts (Figure 1). Its references and citation contexts define the targets, and its title and abstract are used only in a hindsight test (§3.7). A case whose P^+ fails the eligibility checks is dropped rather than replaced by a later paper (Appendix A).

3.2 TARGETS: WHICH WORKS COUNT AS CORRECT

A ranker could score well simply by resurfacing works the researcher already cites, which anticipates nothing. The targets therefore come in three tiers, each removing more of what the researcher could already have known. **CiteNext** contains every work P^+ cites that was published before P. **CiteNew**, the primary tier, removes works the researcher had cited before P. **CiteNew-Isolated** also removes works cited before P by anyone who coauthored with the researcher in the preceding

five years, since such works could have reached the researcher through a collaborator. This progressive design separates genuine anticipation from familiarity with what the researcher already knew (Appendix B.1).

3.3 TARGETS GROUNDED IN EVIDENCE OF INFLUENCE

The tiers above use only bibliography membership. A citation shows that a paper used a work, not that the work shaped it; many references acknowledge background, tools, or contrasts. Two further sources narrow the targets to works that actually shaped the paper, which we call its *roots*: one source asks its authors, the other reads its text. Both use the same three novelty tiers (§3.2), written Author-Root-Next, -New, and -Isolated, and Context-Root-Next, -New, and -Isolated.

What authors say shaped their paper. After P^+ was published, we invited its corresponding author to name the works that shaped or inspired it, offering a shortlist of up to eight references from its bibliography and the option to name any other work; a person matched each of the 1,482 responses to specific papers and reviewed them (Appendix D.1). A case is evaluable with a response, a usable history, and an attested work in the pool: 580 cases for *Author-Root-All* (which also keeps attested works P^+ does not cite), 460 for Author-Root-Next, 359 for Author-Root-New, and 286 for Author-Root-Isolated. These author-collected cases form a distinct set, not restricted to last authors, and are ranked against a 2.60M-paper pool that adds the attested works to the benchmark candidates (Appendix D).

How the next paper uses a work. The second source reads the citation contexts of P^+, each labeled under a six-category taxonomy of citation function; MUSES itself uses only the two-way collapse into *generative* versus not, while the six-category labels of the human-coded contexts are released for other uses (Appendix B.2, Table 10). A cited work enters the Context-Root set when at least one of its contexts is generative. Two annotators agreed at $\kappa = 0.782$ on this binary distinction across 300 contexts. At benchmark scale, a Qwen3-8B classifier (Qwen Team, 2025) with a LoRA adapter (Hu et al., 2022), distilled from a larger teacher, assigns the labels (validation: Appendix B.2). The label describes how the paper uses a work, which can differ from what its authors privately intended.

3.4 ELIGIBILITY, SPLITS, AND THE CANDIDATE POOL

A candidate trajectory (a, P, P^+) is *eligible* when a has at least three papers before P, P^+ appears within five years of P, cites at least ten references, and is a research article. Authors are disjoint across splits, assigned by career midpoint: training before 2018, validation in 2018–2020, and test in 2021–2023, with P dated inside the assigned period; the midpoint sets the split only and is not a retrieval input (Appendix A.3).

The candidate pool is fixed across all cases: it contains every eligible target paper, across splits and tiers, that has a title and abstract, plus 100,000 background papers sampled across publication years, for 2.33 million candidates in total. Each case's targets thus compete with other cases' targets as well as with background papers, so nearly every distractor is a work that some researcher later cited rather than an arbitrary paper. A case is *represented* when at least one of its targets lies in its candidate set $C(t)$, which holds for 134,451 of 140,030 CiteNew test cases and 6,707 of 11,189 Context-Root-New cases (Table 4).

3.5 RETRIEVERS

To see whether more sophisticated representations pay off, the retrievers span the main families of methods, from popularity through lexical, learned-sparse, general-purpose dense, and citation-trained dense models to a supervised model (Appendix C). They map what current families of methods achieve rather than bound the task; MUSES is meant as a target for stronger systems.

Popularity ranks candidates by citations received before P. BM25 (Robertson & Zaragoza, 2009), SPLADE (Formal et al., 2021) (used as released), and a Qwen3-Embedding-8B baseline (Zhang et al., 2025) match candidate titles and abstracts against a *compact query*: the titles, then the abstracts, of the researcher's 20 most recent papers, truncated to 128 words. The remaining methods

embed each history paper separately, then summarize the history either as *1 centroid*, the mean of the history vectors, or as up to *16 centroids* from spherical clustering, which lets separate lines of a researcher's work stay apart. A candidate's score is its highest cosine similarity to any centroid. We evaluate both summaries for GritLM-7B (Muennighoff et al., 2025), SciNCL (Ostendorff et al., 2022), and SPECTER2-Proximity, the SPECTER2 encoder with its proximity adapter (Cohan et al., 2020; Singh et al., 2023), and the 16-centroid summary for Qwen3-Embedding-8B. *Linear-Proximity*, the supervised baseline, learns linear maps for history and candidate SPECTER2-Proximity vectors from training-split citations, selects its checkpoint on validation data, and ranks with 16 centroids; every other model is used frozen. Our *reference retriever*, used in the analyses that need a single ranking, is SPECTER2-Proximity with 16 centroids; it was fixed before results were examined.

3.6 METRICS

Before a project exists, a researcher needs one good entry point into the relevant literature more than a complete list, so the primary metric asks whether any target appears near the top. ***Reach***, or Hit@K, is the share of cases with at least one target in the top K. With E_i the targets of case i that are in its candidate set, $L_i(K)$ its top K results, and I the evaluated cases,

$$\mathrm{Hit@}K = \mathrm{mean}_{i \in I} \mathbb{1}[\,|L_i(K) \cap E_i| > 0\,]. \tag{1}$$

Coverage, or macro Recall@K, is the share of a case's retrievable targets found in the top K, averaged over cases with at least one such target:

$$\mathrm{MacroRecall@}K = \mathrm{mean}_{i \in I,\ |E_i|>0} \frac{|L_i(K) \cap E_i|}{|E_i|}. \tag{2}$$

Coverage and the depth analysis use represented cases; reach is reported over all labeled test cases for the bibliography tiers, where only 4% of cases lack a target in the pool, and over represented cases for the Context-Root and Author-Root targets, where far more do (40% for Context-Root-New), so that retrievers are compared on cases they could reach. Appendix E.2 reports the other variants. We report $K = 100$ throughout, a standard retrieval-evaluation depth, with $K = 10$ and $K = 1000$ as sensitivities.

3.7 THREE DIAGNOSTICS: A PERSONAL-SIGNAL CONTROL, POOLED RANKINGS, AND A HINDSIGHT TEST

Is the signal specific to the researcher, not just their field? Each case is also paired with three *matched histories* from other researchers. These share the researcher's primary field, focal year, and history-size band, share no earlier papers with the researcher, and are the nearest such histories by mean embedding; matching uses only information from before the focal year. We rank each history with base SPECTER2 against the same targets and candidates and compare the focal history's nDCG@1000 with the average over the matched histories (Appendix E.1).

How much of the remaining gap does pooling every ranking close? This second diagnostic pools seven rankings on the 6,707 represented Context-Root-New cases: popularity, BM25, SPLADE, the compact-query Qwen3-Embedding-8B baseline, the reference retriever, GritLM-7B with 16 centroids, and Linear-Proximity. A case counts as reached at depth K if any of them places a target in its top K. This union is an optimistic envelope that inspects up to $7K$ candidates per case, not a combined ranker (Appendix F.1).

How much of the rest closes once the project itself is known? This *hindsight test* keeps the pool, the cutoff, and the reference retriever's history centroids fixed and adds information from P^+ itself: nothing (H), one vector for its title (H+T), or one vector jointly encoding its title and abstract (H+T+A). A candidate's score is its highest similarity to any history or added vector. Because this information does not exist before the project, the test measures dependence on it and is not a usable method.

Table 2: **History beats popularity about six-fold (5.7–6.8× on CiteNew); the choice of retriever moves reach by about six points at most.** Reach (Hit@100, %) of the seven rankings. Bibliography columns use all labeled test cases; Context-Root-New uses represented cases (N = 6,707; 6,704 for history-based methods, which need a usable history); all rank the 2.33M-paper pool. The shaded row is the reference retriever used in later analyses; other representations appear in Appendix Tables 5 and 9. **Bold**: best per column.

			Bibliography			**Citation context**
Family	Retriever	Representation	CiteNext *any citation* N=140,278	CiteNew *new to researcher* 140,030	CiteNew-Isolated *new to coauthors* 139,573	Context-Root-New *generative use* 6,707
Prior only	Popularity	none	6.76	5.01	2.40	2.28
Lexical/sparse	BM25	compact query	32.71	28.68	25.86	13.01
	SPLADE	compact query	32.95	28.83	25.92	13.87
General dense	Qwen3-Embedding-8B	compact query	35.31	31.34	28.39	16.90
	GritLM-7B	16 centroids	**38.91**	**34.25**	**31.06**	**17.63**
Citation dense	SPECTER2-Proximity	16 centroids	36.05	31.17	28.01	15.86
Supervised	Linear-Proximity	16 learned centroids	33.24	28.58	25.45	14.71

4 RESULTS

Across popularity, lexical, sparse, dense, citation-trained, and supervised retrievers, the results support one conclusion: a researcher's past reveals where their next paper will look, but least well what will be new to them. The signal is real and specific to the researcher (§4.1), it fades as works become newer to the researcher (§4.2), and much of the rest depends on the project itself (§4.3).

4.1 PAST WORK ANTICIPATES THE NEXT PAPER'S SOURCES, AND THE SIGNAL IS PERSONAL

Ranking by past work beats ranking by popularity about six-fold. At depth 100, the six history-based rankings in Table 2 find a CiteNew target for 28.58–34.25% of cases, against 5.01% for popularity. History beats popularity on every tier and on the 6,707 represented Context-Root-New cases (13.01–17.63% versus 2.28%). The two kinds of targets rank the systems broadly alike, with the same top three, and the analyses below use the reference retriever, fixed in advance, rather than the best ranking (other representations: Appendix Table 5).

The signal is specific to the researcher. Replacing a researcher's history with three matched histories from other researchers in the same field and year (§3.7) cuts nDCG@1000 to roughly a tenth of its value on all three tiers (Figure 2A; base SPECTER2; about 28,000 test cases with rankings for all four histories; Appendix E.1). The gap is not an artifact of topical distance: choosing the matched histories closest in content leaves it at 89–90% of its size under matching on field, year, and history size alone (Table 25).

One fixed retriever finds a work the authors say shaped their paper in over half of cases. For works that responding authors named as shaping their next paper, including works they had cited before, the reference retriever finds one in its top 100 for 56.03% of the 580 evaluable cases. When the named work is new to the researcher, it does so for 36.21% of 359 cases. A random ranking would do so for well under 1% of cases. Every representation we evaluate reaches 38.28–59.14% of these cases (Appendix Table 21). Author reports also reach beyond the bibliography: in 14 cases, a named work is one that P^+ does not cite (Appendix D.3).

Retrieval finds entry points, not complete reading lists. On the 134,451 represented CiteNew cases, the reference retriever reaches 32.47% at depth 100 but recovers only 9.52% of each case's retrievable targets (Appendix Figure 5; Appendix Table 28).

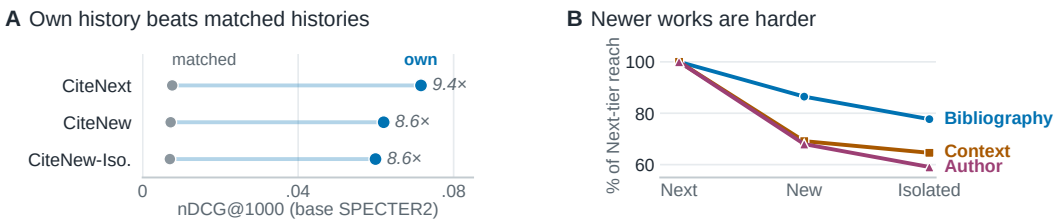

Figure 2: **The signal in a researcher's history is personal, and it fades as works become newer to the researcher.** (A) nDCG@1000 of the researcher's own history versus three matched histories of other researchers (base SPECTER2); labels give the ratio. (B) Hit@100 by novelty tier, indexed to each source's Next-tier value; all three use the reference retriever, with citation-context values from the history-only condition of the hindsight test.

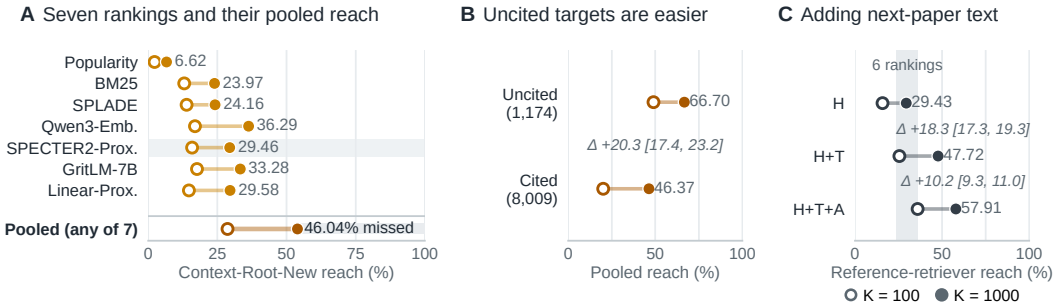

Figure 3: **Nearly half of papers that build on new shaping works stay out of reach; much of what is missing is the project itself.** Represented Context-Root-New cases at $K = 100$ (hollow) and $K = 1000$ (filled). (A) The seven rankings ($N = 6{,}707$; reference retriever shaded) and their pooled reach. (B) Pooled reach, counted per target work, for works never cited before P versus already cited. (C) Hindsight test ($N = 6{,}704$): history (H), plus title (H+T), plus title and abstract (H+T+A); band: six history-based rankings at $K = 1000$. Δ: paired differences, 95% CIs (Appendix F).

4.2 Hard means new to the researcher, not new to the field

In all three kinds of evidence, reach falls as works become newer to the researcher. On the bibliography tiers, every retriever's reach falls from CiteNext to CiteNew to CiteNew-Isolated (the reference retriever: 36.05%, 31.17%, 28.01%). The reference retriever's reach on citation-context and author targets falls the same way (22.85%, 15.80%, 14.76% and 53.26%, 36.21%, 31.47%; Figure 2B), and so does every other representation we evaluate (Appendix Tables 12 and 21); the loss from Next to New is smallest for bibliography (14%) and roughly twice as large for citation context and author evidence (31% and 32%).

For nearly half of the papers that build on new works, no ranking finds one in its top 1,000. Pooling the seven rankings of Table 2, 28.67% of the 6,707 represented Context-Root-New cases are reached at depth 100 and 53.96% at depth 1,000. The remaining 46.04% (3,088 cases) have no target in any ranking's top 1,000 (Figure 3A; Table 31). The misses are not explained by how long or how recent a researcher's history is: pooled reach at depth 1,000 stays within 52.10–55.46% across history-size groups and 52.29–54.78% across recency groups (Appendix F.5). Illustrative misses involve a topic shift (tourism demand to teacher creativity), while reached cases extend an existing line of work, as when a tumor-prognosis history reaches a cancer-mechanism paper (Appendix F.6).

Works no one had cited are found more often than already-cited ones. Of the 9,183 Context-Root-New target works in the pool, the pooled rankings place 66.70% of those with no citations before the cutoff in some top 1,000, against 46.37% of those already cited (Figure 3B). Counting cases instead of works gives the same direction (68.80% versus 51.92%), and so does the reference retriever alone at both depths (Appendix F.7). History-based retrieval surfaces work the field has not yet noticed more readily than distant, well-known work.

4.3 Much of what is missing is the project itself

A title's worth of project text helps more than any change of retriever we tested. In the hindsight test (§3.7), on 6,704 matched cases, reach at depth 1,000 rises from 29.43% with history alone to 47.72% when the next paper's title is added (+18.29 points; 95% CI $[17.33, 19.26]$, resampling authors). It rises further to 57.91% when a joint encoding of the title and abstract replaces the title vector (+10.19; $[9.34, 11.02]$; Figure 3C). Excluding 258 cases in which the next paper's text names a target leaves the title effect essentially unchanged (+18.27; Appendix F.4).

The limit is information more than representation. Swapping retrievers is worth far less than the project's title. At depth 100, the six history-based rankings fall within 4.6 points of one another on Context-Root-New (13.01–17.63%), whereas adding the next paper's title to the reference retriever's history adds 9.69 points (15.80% to 25.49%). Neither citation pretraining nor supervision on MUSES beats the best general-purpose embedder: GritLM-7B has the highest reach in every column of Table 2, though on Context-Root-New its lead over Qwen3-Embedding-8B is under a point (17.63% versus 16.90%), and Linear-Proximity, trained on MUSES citations, falls below the frozen encoder it starts from (28.58% vs. 31.17% on CiteNew).

How a history is summarized matters for the works authors name. Up to 16 centroids rather than one mean vector raise Author-Root-All reach for all four encoder families (+3.6 to +15.3 points, paired 95% CIs above zero; Appendix Table 22); on the bibliography tiers they help at depth 10, though GritLM's mean vector is stronger deeper (Appendix Table 5).

5 Discussion and Limitations

Implications for research agents and retrievers. Systems that surface literature before a project exists can act on four results. *Condition on the individual:* matched histories from the same field keep only about a tenth of the signal (§4.1). *Target what is new to the researcher, not what is obscure:* reach falls with newness to the researcher, whereas targets without prior citations are reached more often (§4.2). *Ask early:* reach nearly doubles once the next paper's title and abstract are known (§4.3), so even a working title could recover much of the gap. *Model interest change:* the misses we inspected are topic shifts (§4.2), suggesting the open problem is anticipating where interests move.

Limitations and what would resolve them. *Is a classifier label trustworthy?* Its precision is 0.63 on the human-coded development contexts used to select it; a held-out sample would measure accuracy directly. *Are the author reports representative?* They come from authors who chose to respond, and the bibliography-derived shortlist may anchor their answers (Appendix D.1). *Could models have seen the answers?* GritLM-7B and Qwen3-Embedding-8B postdate the test period and cannot be audited, but they only score text similarity, and BM25, which learns nothing, reproduces the tier decline and the unreached share (Appendix G). A generative LLM, which may have memorized the next paper and its references, cannot be made time-safe, so we do not prompt one. *Would better retrievers close the gap?* Possibly: the seven rankings are neither exhaustive nor optimized, so the residual describes them, not the limit of the task. *Does the last-author convention hold everywhere?* Less so where authors are ordered alphabetically, as often in mathematics and economics, and histories inherit S2ORC's author-disambiguation errors. *Does the candidate pool limit the conclusions?* It holds the roughly half of CiteNew targets with a title and abstract; broader pools would test the patterns beyond it. *What about influence that leaves no citation?* The tiers remove only recorded citations, and citations and author reports observe only part of influence; elicitation not anchored to the bibliography could reach conversations and uncited reading. *Which signals are left out?* Retrievers see only titles and abstracts, not full text or visibility such as preprints and talks; widening the inputs would show how much of the residual is recoverable before the project. *Do entry points help researchers?* Appendix D.4 shows examples, but only a user study can measure the benefit.

Conclusion. A researcher's past reveals where their next paper will look, but least well what will be new to them, and much of the rest depends on the project itself. By measuring where that signal stops, MUSES lets future systems be judged by how much of the rest they recover as a project takes shape.

ETHICS STATEMENT

MUSES is built from public bibliographic records. Author responses were collected with consent for research use, and the release includes only identifier-level structured data, excluding raw responses, free text, and contact details (Appendix I). Retrieval conditioned on an individual researcher's history could be misused to profile researchers or rank scientific influence; MUSES evaluates retrieval of literature for a researcher's own use.

REPRODUCIBILITY STATEMENT

Appendix A specifies corpus construction, splits, and case sets; Appendix B the citation tiers and Context-Root-New labels; Appendix C every retrieval system; Appendices D–F the author-collected cases, matched histories, the seven-ranking results, and the hindsight test; Appendix G the encoder-exposure checks; Appendix H three supporting checks run on an earlier version of the benchmark; and Appendix I the identifier-only release, which reproduces every metric on the final benchmark from saved rankings without redistributing S2ORC text or author responses.

AI USE STATEMENT

W used generative AI tools to edit the paper for readability, to create and edit code and figures. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

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/.

Hiba Arnaout, Noy Sternlicht, Tom Hope, and Iryna Gurevych. In-depth research impact summarization through fine-grained temporal citation analysis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6749–6789, 2026.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, et al. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*, 2026.

Peter Brown, RELISH Consortium, and Yaoqi Zhou. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019:baz085, 2019.

Eric Chamoun, Yizhou Chi, Yulong Chen, Rui Cao, Zifeng Ding, Michalis Korakakis, and Andreas Vlachos. SciPaths: Forecasting pathways to scientific discovery. *arXiv preprint arXiv:2605.14600*, 2026.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3586–3596, 2019.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2270–2282, 2020.

Travis Ebesu and Yi Fang. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 1093–1096, 2017.

Markus Flicke, Glenn Angrabeit, Madhav Iyengar, Vitalii Protsenko, Illia Shakun, Jovan Cicvaric, Bora Kargi, Haoyu He, Lukas Schuler, Lewin Scholz, Kavyanjali Agnihotri, Yong Cao, and Andreas Geiger. Scholar inbox: Personalized paper recommendations for scientists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 307–317, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-demo.30. URL https://aclanthology.org/2025.acl-demo.30/.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292, 2021.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 421–430, 2010.

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.

Chi Jiang, Xiao Ma, Jiangfeng Zeng, Yin Zhang, Tingting Yang, and Qiumiao Deng. TAPRec: Time-aware paper recommendation via the modeling of researchers' dynamic preferences. *Scientometrics*, 128(6):3453–3471, 2023. doi: 10.1007/s11192-023-04731-4. URL https://doi.org/10.1007/s11192-023-04731-4.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 2026. URL https://aclanthology.org/2026.findings-acl.644/.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4969–4983, 2020.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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, 2022.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*, 2024.

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

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, 2023.

Neil R Smalheiser. Literature-based discovery: Beyond the abcs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224, 2012.

Noy Sternlicht and Tom Hope. CHIMERA: A knowledge base of scientific idea recombinations for research analysis and ideation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1871–1905, 2026.

Kazunari Sugiyama and Min-Yen Kan. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 29–38, 2010. doi: 10.1145/1816123.1816129. URL https://doi.org/10.1145/1816123.1816129.

Don R Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*, 2015.

Daniel Vodička, Pavel Kral, Christophe Cerisara, and Jakub Šmíd. Large language models for citation function classification. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 2430–2439, Palma de Mallorca, Spain, 2026. ELRA Language Resource Association. doi: 10.63317/4sb25z5kxz3q. URL https://aclanthology.org/2026.lrec-1.191/.

Simon Wakeling, Monica Lestari Paramita, and Stephen Pinfield. How do authors perceive the way their work is cited? findings from a large-scale survey on quotation accuracy. *Journal of the Association for Information Science and Technology*, 76(10):1396–1410, 2025. doi: 10.1002/asi.70000. URL https://doi.org/10.1002/asi.70000.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *ACL*, 2024.

Yang Zhang and Qiang Ma. On-the-fly citation recommendation based on content-dependent embeddings. *Knowledge-Based Systems*, 343:116052, 2026. doi: 10.1016/j.knosys.2026.116052. URL https://doi.org/10.1016/j.knosys.2026.116052.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Leqi Zheng, Jiajun Zhang, Canzhi Chen, Chaokun Wang, Hongwei Li, Yuying Li, Yaoxin Mao, Shannan Yan, Zixin Song, Zhiyuan Feng, Zhaolu Kang, Zirong Chen, Hang Zhang, Qiang Liu, Liang Wang, and Ziyang Liu. What should i cite? a RAG benchmark for academic citation prediction. In *Proceedings of the ACM Web Conference 2026*, 2026. doi: 10.1145/3774904.3792075. URL https://arxiv.org/abs/2601.14949.

APPENDIX GUIDE

Use this guide to locate the definitions and supporting results behind the main paper.

Benchmark construction and case sets. Corpus, case construction, splits, and the number of cases behind each result: Appendix A, especially Table 4.

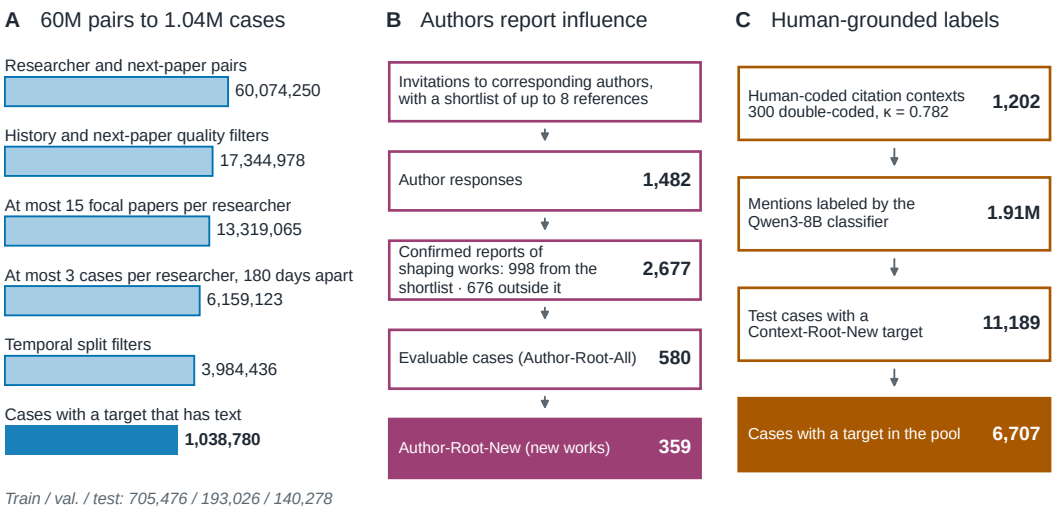

Figure 4: **How MUSES was built.** (A) Benchmark construction from S2ORC; each stage is nested in the previous one, and bar length is log-scaled (Table 3). (B) Author attestation: invitations carried a shortlist of up to eight references from the next paper's bibliography, and authors could also name unlisted works (Appendix D.1). (C) Citation-context labels: human-coded contexts ground the classifier that labels all mentions (Appendix B.2). Stages in B and C are process steps, not nested counts; one response can yield several reports.

Targets, labels, and retrievers. Bibliography tiers and Context-Root labels: Appendix B; retrieval systems and how each maps to the result tables: Appendix C, Tables 13 and 14; how author reports become evaluable cases: Appendix D, Tables 18–19.

Reach and coverage. Represented cases, Hit@K, macro Recall@K, and depth: Appendix E.2, Tables 26 and 28; alternative denominators: Table 29.

Matched histories and author results. Construction of matched histories and their paired effects: Appendix E.1; full author results: Appendix D.2.

Context-Root-New retrieval and the hindsight test. Appendix F: the seven rankings and their pooled reach (Table 31), targets outside the pool (Table 33), the hindsight test (Table 34), and reach by history and target profile (Tables 35 and 39).

Checks on an earlier version. A pre-2021 encoder, targets matched on count, and the number of centroids: Appendix H.

Release and reproduction. What is released, how to reproduce the results, privacy, and licensing: Appendix I.

A Benchmark construction and temporal eligibility

This section defines a benchmark case, its cutoff, and the case sets used in evaluation. Figure 4 summarizes how the benchmark cases and the two human-grounded evidence layers were built.

A.1 Corpus, task unit, and release construction

The benchmark uses paper, authorship, reference, publication-date, title, and abstract records from the Semantic Scholar Open Research Corpus (S2ORC), release 2026-03-10. An evaluation case is an author–focal-paper pair. For an eligible author, papers are ordered by publication date within the last-author trajectory. The next paper P^+ is the immediately subsequent last-author paper after focal P; the case is retained only if this paper passes the reference-count, paper-type, and maximum-gap eligibility filters. Construction does not skip an ineligible P^+ to search for a later replacement. The

Table 3: **Construction keeps 1.04M of 60.1M role-consistent focal/next-paper candidates, split author-disjointly.** Stages are nested; each applies the filter named in its row.

Construction stage	Instances or candidates
Role-consistent focal / next-paper candidates	60,074,250
After pre-history and next-paper quality filters	17,344,978
After at most 15 focal papers per author	13,319,065
After at most 3 cases per author, 180 days apart	6,159,123
After temporal split filters	3,984,436
After requiring a target with title and abstract	1,038,780
Train / validation / test	705,476 / 193,026 / 140,278

author history contains that author's papers published strictly before P; P itself is excluded from the history query. Each case is ranked against a shared candidate pool.

Construction first requires usable paper dates and bibliographic links, then applies the author-role, history, next-paper, and target-quality rules: at least three prior papers, an immediately subsequent last-author paper within five years, at least ten references in that paper, and exclusion of review, meta-analysis, news, editorial, letters/comments, dataset, and book records. It then assigns splits by author career midpoint, samples at most 15 candidate focal papers per author, keeps at most three cases per author with at least 180 days between them, and finally keeps cases with at least one target that has a title and abstract. Retrieval candidates come from the shared fixed pool of 2,330,779 papers: the union of all 2,230,779 benchmark target papers across splits and citation tiers that have title and abstract text, plus 100,000 background papers with title and abstract sampled with year stratification (seed 42); systems that use text require title and abstract text. A candidate is eligible when its publication date is at least one day before the focal publication date. Full-text availability is not required. All eligibility, sampling, and query-construction thresholds in this section (paper counts, gap and reference-count minimums, sampling caps, and spacing) were fixed during benchmark construction, before any retrieval evaluation was run, and are not tuned to reported outcomes; they balance benchmark scale, per-author diversity, and text availability rather than any evaluated metric.

The benchmark contains 1,038,780 cases: 705,476 train, 193,026 validation, and 140,278 test cases. The test split is author-disjoint from training and covers focal dates from 2021 through 2023. The next-paper target must satisfy the reference-count, paper-type, and maximum-gap filters; a case without a usable next-paper bibliography is excluded. The construction funnel and final split counts are in Table 3.

Each case records the researcher, P, P^+, the cutoff date, its split, and its targets in each tier; Appendix I describes the released files. The Context-Root and retrieval results use subsets of the test split, each defined where it is reported and summarized in Table 4.

A.2 Observable information and temporal candidate rule

For focal paper P with date t, the candidate set is

$$C(t) = \{d \in \mathcal{G} : \mathrm{pubdate}(d) < t\},$$

where $\mathcal{G}$ is the fixed shared candidate pool. Every ranked candidate belongs to the shared pool and is dated strictly before P.

Retrievers may use the researcher's papers published strictly before P and the eligible candidates, including their text when the method uses text. P itself is excluded from the history because it sets the cutoff. The next paper P^+ supplies the outcome: its bibliography and citation contexts define the targets but are never retrieval inputs. Context-Root labels require a citation context, so a bibliography target without a recoverable context is left unlabeled rather than counted as non-generative. The title and abstract of P^+ enter only the hindsight test.

A future target is unavailable when it is absent from the shared pool or published after the focal cutoff. Unavailable targets are excluded from represented-case results rather than counted as ranking misses. In the three Context-Root tiers, every unavailable target is absent from the pool; none

Table 4: **Test-split case sets behind each result.** For the two Context-Root rows, the columns give the corresponding Context-Root tier (Context-Root-Next, -New, -Isolated).

Case set	CiteNext	CiteNew	CiteNew-Isolated
All labeled test cases	140,278	140,030	139,573
Represented test cases	135,211	134,451	133,373
Cases with a Context-Root target (in or out of pool)	12,786	11,189	9,624
Represented Context-Root cases	7,851	6,707	5,735

is unavailable because of a post-cutoff publication date. Availability is computed separately for bibliography and Context-Root targets.

A.3 Split policy and populations

Cases are split by author using each author's career midpoint: training cases before 2018, validation cases in 2018–2020, and test cases in 2021–2023. Different results use different case sets, and Table 4 lists them. *All labeled* cases have at least one target in the tier. *Represented* cases have at least one target in the candidate set $C(t)$. For the Context-Root tiers, a case must first have at least one classifier-labeled Context-Root target, and is represented when one such target is in $C(t)$.

Of the 11,189 test cases with a Context-Root-New target, 4,482 have none in the pool; they cannot be ranked and are not counted as misses. Appendix B.2 defines the Context-Root label.

B Bibliography tiers and Context-Root labels

This section defines the bibliography tiers and the citation-context labels behind the Context-Root tiers.

B.1 Bibliography tiers

Let $Y_{\text{next}}(P)$ be the set of distinct papers published before P and cited by its immediately subsequent last-author paper P^+, provided that P^+ passes the case-retention filters. The bibliography tiers are defined on the same target bibliography:

$$Y_{\text{CiteNext}}(P) = Y_{\text{next}}(P),$$
$$Y_{\text{CiteNew}}(P) = Y_{\text{next}}(P) \setminus B_a(t),$$
$$Y_{\text{CiteNew}-\text{Isolated}}(P) = Y_{\text{CiteNew}}(P) \setminus S_a(t),$$

where t is the date of P, $B_a(t)$ contains the papers cited by researcher a before P (their prior bibliography), and $S_a(t)$ contains the papers cited before P by anyone who coauthored any of a's papers in the five years before P. The five-year window selects recent coauthor relationships; that coauthor's citation may be older than five years, and the coauthor need not have appeared on P. Coauthors who first collaborate with the researcher on P^+ fall outside the window, so their prior citations remain unmasked. Thus CiteNew-Isolated $\subseteq$ CiteNew $\subseteq$ CiteNext. In the test split, the recent-coauthor filter removes 12.96% of CiteNew target pairs (284,785 of 2,197,383; Table 27): most works new to the researcher had not been cited by a recent coauthor, but about one in eight had. These are recorded author/coauthor citation-history exclusions; they do not establish first exposure or author intent. The released records also list targets from the researcher's next three papers, but every evaluation here uses only the immediately following paper.

Full bibliography-tier results. The tables below report Hit@K (reach), nDCG@K, observed-target Recall@K, and MRR on all labeled test cases at three depths; a case without a ranking stays in the denominator and scores zero. In the row labels, Proximity means SPECTER2-Proximity, GritLM means GritLM-7B, and Linear-Proximity means the supervised linear-set SPECTER2-Proximity method. The tables include both 1-centroid and 16-centroid history representations for each family. MRR is reported for completeness.

In these tables, a relevant target is a distinct paper in the named tier, and Y_i denotes all of case i's targets in that tier, including those not in the pool. Hit@K follows Eq. (1). Observed-target

Table 5: **Every system's reach falls from CiteNext to CiteNew to CiteNew-Isolated at every depth.** Bibliography-tier Hit@K (percent) on all labeled test cases.

Tier	Retrieval system	N	@10	@100	@1000
CiteNext	Proximity, 16 centroids	140,278	26.46%	36.05%	52.76%
CiteNext	Proximity, 1 centroid	140,278	18.36%	31.21%	49.75%
CiteNext	GritLM, 16 centroids	140,278	28.15%	38.91%	56.76%
CiteNext	GritLM, 1 centroid	140,278	26.83%	40.13%	59.20%
CiteNext	SciNCL, 16 centroids	140,278	25.52%	34.07%	50.04%
CiteNext	SciNCL, 1 centroid	140,278	16.62%	29.80%	49.23%
CiteNext	Linear-Proximity	140,278	24.21%	33.24%	50.16%
CiteNew	Proximity, 16 centroids	140,030	20.89%	31.17%	49.85%
CiteNew	Proximity, 1 centroid	140,030	13.87%	26.94%	46.94%
CiteNew	GritLM, 16 centroids	140,030	22.27%	34.25%	54.24%
CiteNew	GritLM, 1 centroid	140,030	21.41%	36.03%	57.01%
CiteNew	SciNCL, 16 centroids	140,030	20.15%	29.25%	47.05%
CiteNew	SciNCL, 1 centroid	140,030	12.30%	25.64%	46.45%
CiteNew	Linear-Proximity	140,030	19.10%	28.58%	47.12%
CiteNew-Isolated	Proximity, 16 centroids	139,573	18.23%	28.01%	46.79%
CiteNew-Isolated	Proximity, 1 centroid	139,573	11.91%	23.89%	43.85%
CiteNew-Isolated	GritLM, 16 centroids	139,573	19.52%	31.06%	51.49%
CiteNew-Isolated	GritLM, 1 centroid	139,573	18.68%	32.70%	54.35%
CiteNew-Isolated	SciNCL, 16 centroids	139,573	17.53%	26.12%	44.15%
CiteNew-Isolated	SciNCL, 1 centroid	139,573	10.46%	22.67%	43.51%
CiteNew-Isolated	Linear-Proximity	139,573	16.64%	25.45%	44.07%

Recall@K is $|L_i(K) \cap Y_i|/|Y_i|$, averaged over cases; unlike the coverage of Eq. (2), it divides by all of a case's targets rather than only those in the pool, so it is lower (4.90% versus 9.52% for the reference retriever on CiteNew at $K = 100$; Tables 28 and 29). Binary nDCG@K is

$$\mathrm{nDCG@}K_i = \frac{\sum_{r=1}^{K} \mathbb{1}[\ell_{i,r} \in Y_i] / \log_2(r+1)}{\sum_{r=1}^{\min(K,|Y_i|)} 1/\log_2(r+1)},$$

where $\ell_{i,r}$ is the candidate at rank r; the base of the logarithm cancels. Because the ideal ranking in the denominator places every target, including those not in the pool, absolute nDCG values are low. Tables 5–8 divide by all labeled test cases, including cases with no target in the pool or no ranking, so they are slightly lower than the represented-case values in Appendix E.2. With about 140,000 cases per tier, sampling intervals are a fraction of a point and are not shown.

Other retrievers on the bibliography tiers. Popularity, BM25, SPLADE, the compact-query Qwen3-Embedding-8B baseline, and Qwen3-Embedding-8B with 16 centroids are reported at Hit@100, the depth available for all of them.

B.2 Context-Root labels

This section specifies what a Context-Root label claims, how the labels were produced, and what the classifier's validation does and does not establish. The label concerns the role a cited work plays in a citation context: whether the citing paper uses it as a generative input to its own ideas. It is separate from an author's report of inspiration. The six-category taxonomy distinguishes three generative functions from three other uses (Table 10). A cited work is a Context-Root target of P^+ when at least one of its citation contexts in P^+ is labeled generative; a cited work with no recoverable context stays in the bibliography tiers but receives no Context-Root label. The main results require at least one generative context; a stricter rule requiring at least two is reported below as a check.

Table 6: **Bibliography-tier nDCG@K (percent) on all labeled test cases.**

Tier	Retrieval system	N	@10	@100	@1000
CiteNext	Proximity, 16 centroids	140,278	7.34%	6.28%	8.17%
CiteNext	Proximity, 1 centroid	140,278	4.52%	4.38%	6.21%
CiteNext	GritLM, 16 centroids	140,278	7.93%	7.08%	9.32%
CiteNext	GritLM, 1 centroid	140,278	7.56%	7.01%	9.37%
CiteNext	SciNCL, 16 centroids	140,278	7.00%	5.91%	7.77%
CiteNext	SciNCL, 1 centroid	140,278	3.93%	4.00%	5.96%
CiteNext	Linear-Proximity	140,278	6.48%	5.48%	7.26%
CiteNew	Proximity, 16 centroids	140,030	4.86%	4.84%	6.58%
CiteNew	Proximity, 1 centroid	140,030	2.90%	3.34%	5.02%
CiteNew	GritLM, 16 centroids	140,030	5.03%	5.32%	7.43%
CiteNew	GritLM, 1 centroid	140,030	4.92%	5.44%	7.69%
CiteNew	SciNCL, 16 centroids	140,030	4.66%	4.56%	6.27%
CiteNew	SciNCL, 1 centroid	140,030	2.45%	3.02%	4.82%
CiteNew	Linear-Proximity	140,030	4.40%	4.31%	5.93%
CiteNew-Isolated	Proximity, 16 centroids	139,573	4.34%	4.60%	6.23%
CiteNew-Isolated	Proximity, 1 centroid	139,573	2.57%	3.14%	4.70%
CiteNew-Isolated	GritLM, 16 centroids	139,573	4.48%	5.03%	7.01%
CiteNew-Isolated	GritLM, 1 centroid	139,573	4.36%	5.12%	7.24%
CiteNew-Isolated	SciNCL, 16 centroids	139,573	4.17%	4.34%	5.93%
CiteNew-Isolated	SciNCL, 1 centroid	139,573	2.16%	2.83%	4.51%
CiteNew-Isolated	Linear-Proximity	139,573	3.95%	4.12%	5.63%

Labels were assigned from the citation-context text alone, without reference to any retrieval result. The classifier is a two-class Qwen3-8B model with a LoRA adapter, distilled from a GPT-5.4-mini teacher (agreement with the teacher: $\kappa = 0.771$). Each context was truncated to 512 tokens and labeled generative when the predicted probability exceeded 0.5. The classifier labeled 1,912,610 unique citation mentions; 26,399 mentions and 21,633 pairs of next paper and cited work were labeled generative, and 3,335 pairs had at least two generative contexts. These counts describe the labeled corpus, not the prevalence of generative citation in scientific writing generally. The release for this version includes the labels and the classifier adapter (Appendix I).

Context-Root-New is the subset of CiteNew targets that the classifier labels generative. Agreement between the human annotators supports the reliability of the taxonomy, not the accuracy of the classifier.

Two annotators independently labeled a shared sample without seeing classifier outputs. On 300 double-coded contexts, Cohen's κ was 0.782 for generative versus not and 0.715 for the six categories. The 1,202 human-coded development contexts were labeled once and then revised after review; they were also used to choose the classifier configuration, so they are not an independent test set. Against the final (revised) labels, the classifier achieved precision 0.6345, recall 0.9321, and F1 0.7550; against the initial labels, precision 0.7185, recall 0.8906, and F1 0.7953. Only three development mentions also appear among the labeled test-split mentions, so these figures do not directly measure accuracy on the benchmark's labels or at the level of cited works. Context-Root targets are therefore classifier-assigned, not human-verified. A cited work without a recoverable context is unlabeled, not non-generative.

Context-Root tiers and cases. Context-Root-Next, Context-Root-New, and Context-Root-Isolated keep the generative-labeled targets of CiteNext, CiteNew, and CiteNew-Isolated. A case is represented when at least one of its targets in the tier is in its candidate set. The represented sets contain 7,851 Context-Root-Next, 6,707 Context-Root-New, and 5,735 Context-Root-Isolated

Table 7: **Bibliography-tier observed-target Recall@K (percent) on all labeled test cases.** The denominator includes targets not in the pool, so values are lower than the coverage of Eq. (2).

Tier	Retrieval system	N	@10	@100	@1000
CiteNext	Proximity, 16 centroids	140,278	3.01%	5.61%	11.78%
CiteNext	Proximity, 1 centroid	140,278	2.00%	4.60%	10.55%
CiteNext	GritLM, 16 centroids	140,278	3.47%	6.84%	14.26%
CiteNext	GritLM, 1 centroid	140,278	3.33%	7.09%	14.90%
CiteNext	SciNCL, 16 centroids	140,278	2.83%	5.15%	11.17%
CiteNext	SciNCL, 1 centroid	140,278	1.76%	4.45%	10.84%
CiteNext	Linear-Proximity	140,278	2.59%	4.77%	10.56%
CiteNew	Proximity, 16 centroids	140,030	2.57%	4.90%	11.01%
CiteNew	Proximity, 1 centroid	140,030	1.70%	4.09%	9.97%
CiteNew	GritLM, 16 centroids	140,030	2.92%	6.00%	13.48%
CiteNew	GritLM, 1 centroid	140,030	2.83%	6.37%	14.28%
CiteNew	SciNCL, 16 centroids	140,030	2.44%	4.51%	10.48%
CiteNew	SciNCL, 1 centroid	140,030	1.48%	3.95%	10.27%
CiteNew	Linear-Proximity	140,030	2.27%	4.23%	9.90%
CiteNew-Isolated	Proximity, 16 centroids	139,573	2.69%	5.06%	11.25%
CiteNew-Isolated	Proximity, 1 centroid	139,573	1.76%	4.17%	10.16%
CiteNew-Isolated	GritLM, 16 centroids	139,573	3.04%	6.15%	13.76%
CiteNew-Isolated	GritLM, 1 centroid	139,573	2.91%	6.48%	14.57%
CiteNew-Isolated	SciNCL, 16 centroids	139,573	2.56%	4.64%	10.71%
CiteNew-Isolated	SciNCL, 1 centroid	139,573	1.53%	4.02%	10.48%
CiteNew-Isolated	Linear-Proximity	139,573	2.38%	4.36%	10.14%

cases. Hit@100 is the share of these cases with a Context-Root target in the first 100 results. The comparison below scores bibliography and Context-Root targets on the same cases and the same ranking; it compares two kinds of targets, not an intrinsic difficulty.

Of the Context-Root target pairs, 11,435 / 9,183 / 7,476 (Context-Root-Next / -New / -Isolated) are in the pool; the remaining 10,198 / 8,412 / 6,889 are not. None is unavailable because of a post-cutoff publication date. Among the 1,664 represented Context-Root-New cases with at least two targets in the pool, the reference retriever's rate of finding both at $K = 100$ exceeds the rate expected under within-case independence of the two targets' marginal hit probabilities by 1.43 percentage points (95% case-bootstrap CI [0.93, 2.01]); targets in the same case are therefore weakly but detectably correlated rather than independent.

For Context-Root-New, the paired bibliography-minus-Context-Root Hit@100 difference is 28.78 percentage points (95% paired bootstrap interval 27.72 to 29.86; 10,000 resamples). Because the two kinds of targets differ in number and composition, the lower Context-Root-New reach does not by itself establish that these works are harder to retrieve; a check with target counts held equal, run on an earlier version of the benchmark, is in Appendix H. Under the stricter rule requiring at least two generative contexts, the represented sets shrink to 1,471 / 1,175 / 937 cases (Context-Root-Next / -New / -Isolated), with Hit@100 of 24.81% / 18.21% / 17.82%; this is a check on the labeling rule, not an alternative main analysis.

Which ranking each result uses. The comparison above and the matched-history analysis use base SPECTER2 with 16 history centroids. The seven-ranking results, the author results, and the CiteNew coverage analysis use the reference retriever (SPECTER2-Proximity, 16 centroids); the hindsight test recomputes that ranking under its own procedure. Numbers from different rankings are therefore close but not identical.

Table 8: **Bibliography-tier mean reciprocal rank (MRR) on all labeled test cases.**

Tier	Retrieval system	N	MRR
CiteNext	Proximity, 16 centroids	140,278	0.204
CiteNext	Proximity, 1 centroid	140,278	0.126
CiteNext	GritLM, 16 centroids	140,278	0.208
CiteNext	GritLM, 1 centroid	140,278	0.199
CiteNext	SciNCL, 16 centroids	140,278	0.199
CiteNext	SciNCL, 1 centroid	140,278	0.111
CiteNext	Linear-Proximity	140,278	0.189
CiteNew	Proximity, 16 centroids	140,030	0.149
CiteNew	Proximity, 1 centroid	140,030	0.087
CiteNew	GritLM, 16 centroids	140,030	0.144
CiteNew	GritLM, 1 centroid	140,030	0.143
CiteNew	SciNCL, 16 centroids	140,030	0.145
CiteNew	SciNCL, 1 centroid	140,030	0.074
CiteNew	Linear-Proximity	140,030	0.139
CiteNew-Isolated	Proximity, 16 centroids	139,573	0.129
CiteNew-Isolated	Proximity, 1 centroid	139,573	0.075
CiteNew-Isolated	GritLM, 16 centroids	139,573	0.125
CiteNew-Isolated	GritLM, 1 centroid	139,573	0.124
CiteNew-Isolated	SciNCL, 16 centroids	139,573	0.125
CiteNew-Isolated	SciNCL, 1 centroid	139,573	0.063
CiteNew-Isolated	Linear-Proximity	139,573	0.120

Table 9: **Popularity, BM25, SPLADE, and Qwen3-Embedding-8B also lose reach from CiteNext to CiteNew to CiteNew-Isolated.** Hit@100 on all labeled test cases in each tier.

Tier	System	N	Hit@100
CiteNext	Popularity	140,278	6.76%
CiteNew	Popularity	140,030	5.01%
CiteNew-Isolated	Popularity	139,573	2.40%
CiteNext	BM25	140,278	32.71%
CiteNew	BM25	140,030	28.68%
CiteNew-Isolated	BM25	139,573	25.86%
CiteNext	SPLADE	140,278	32.95%
CiteNew	SPLADE	140,030	28.83%
CiteNew-Isolated	SPLADE	139,573	25.92%
CiteNext	Qwen3-Embedding-8B (compact query)	140,278	35.31%
CiteNew	Qwen3-Embedding-8B (compact query)	140,030	31.34%
CiteNew-Isolated	Qwen3-Embedding-8B (compact query)	139,573	28.39%
CiteNext	Qwen3-Embedding-8B, 16 centroids	140,278	37.82%
CiteNew	Qwen3-Embedding-8B, 16 centroids	140,030	33.04%
CiteNew-Isolated	Qwen3-Embedding-8B, 16 centroids	139,573	29.87%

Table 10: **Six citation-context roles.** The three generative-use categories form the Context-Root label; the other three are not generative.

Category	Reader-facing definition
Theoretical foundation (TF)	The citing paper explicitly operates within or operationalizes the cited theory, framework, or formalism.
Method extension (ME)	The citing paper explicitly modifies, extends, generalizes, re-derives, or structurally repurposes the cited method.
Generative motivation (GM)	A gap, limitation, finding, suggestion, assumption, or inspiration in the cited work motivates the present paper.
Contrast or comparison (CC)	The citation supports rhetorical contrast or benchmarking without generative uptake.
Tool or resource (TR)	The cited work is used instrumentally as software, data, a benchmark, or a method applied as-is.
Background context (BC)	The citation supplies contextual attribution or evidence without an explicit generative or instrumental role.

Table 11: **Context-Root targets per case and availability by tier.** The mean is over all cases with a Context-Root target in the tier; unavailable targets are absent from the pool.

Tier	Cases with a target	Represented cases	Mean targets per case
Context-Root-Next	12,786	7,851	1.69
Context-Root-New	11,189	6,707	1.57
Context-Root-Isolated	9,624	5,735	1.49

C Retrieval systems and implementation details

This section lists each retrieval system's inputs and scoring. Table 13 gives an overview; Table 15 specifies training and scoring for Linear-Proximity, the supervised method, whose name distinguishes it from base SPECTER2 and from the untrained SPECTER2-Proximity encoder.

Compact query. The query takes up to 20 of the researcher's most recent papers with text, in any author position, published strictly before P (excluding P). It joins their titles, most recent first, followed by their abstracts in the same order, normalizes whitespace, and truncates the result to 128 words. Candidate text is the title followed by the first 1,500 characters of the abstract. BM25, SPLADE, and the compact-query Qwen3-Embedding-8B baseline use this same query.

Number of centroids. The number of history centroids is fixed at up to 16 rather than tuned; on an earlier version of the benchmark, reach changed by less than a point beyond 8 centroids (Appendix H).

1-centroid history. The 1-centroid variants average the vectors of all of the researcher's papers before P into one L2-normalized vector and rank candidates by cosine similarity; Linear-Proximity applies the same averaging after its learned history map.

Which ranking each result uses. Three SPECTER2-based rankings appear in the paper and are distinct: the reference retriever (SPECTER2-Proximity, 16 centroids), used in the main tables, the author results, and the seven-ranking results; base SPECTER2 with 16 centroids, used for the matched-history analysis and the same-case bibliography-versus-Context-Root comparison; and the history-only condition H of the hindsight test, which recomputes the reference retriever under the hindsight procedure (Table 14). Likewise, the compact-query Qwen3-Embedding-8B baseline differs from the 16-centroid Qwen3-Embedding-8B variant, while BM25 uses one configuration throughout.

For some runs, the records do not preserve the exact model checkpoint, tokenizer or index settings, ranking-file checksum, or software revision. We do not guess these values; the release lists them as unavailable, and the saved rankings reproduce every reported metric regardless (Appendix I).

Table 12: **On the same cases and ranking, Context-Root targets are reached far less often than bibliography targets.** Hit@100 for the Context-Root tier and for the matching bibliography tier on the same represented cases, base SPECTER2 with 16 centroids. The contrast also reflects the smaller number of Context-Root targets per case (Table 11), so it does not by itself isolate retrieval difficulty.

Tier	Context-Root Hit@100 (n/N)	Bibliography Hit@100, same cases (n/N)
Context-Root-Next	1,841/7,851 (23.45%)	4,405/7,851 (56.11%)
Context-Root-New	1,095/6,707 (16.33%)	3,025/6,707 (45.10%)
Context-Root-Isolated	875/5,735 (15.26%)	2,316/5,735 (40.38%)

Table 13: **Retrieval systems, their inputs, and where each is reported.** Every input is available before the focal paper P.

Method	Input	Scoring and where reported
Popularity	Citation counts of candidates before P; no history	Ranks candidates by citations received before P. Table 2 and the seven-ranking results.
BM25	Compact query and candidate text	Lexical matching. Table 2 and the seven-ranking results.
SPLADE	Same compact query	Learned sparse matching, used as released. Table 2 and the seven-ranking results.
Qwen3-Embedding-8B (compact query)	Same compact query as one vector	Dense similarity. Table 2 and the seven-ranking results; a 16-centroid variant appears in Table 9.
SPECTER2-Proximity	History and candidate vectors from SPECTER2 with its proximity adapter; 1 or 16 centroids	Maximum cosine similarity to the centroids. With 16 centroids it is the reference retriever: bibliography tiers, author results, and the seven-ranking results.
Base SPECTER2	Vectors from the base SPECTER2 encoder; 16 centroids	Used only for the matched-history analysis and the same-case bibliography-versus-Context-Root comparison (Table 12).
GritLM-7B	History and candidate vectors; 1 or 16 centroids	Dense similarity. Bibliography tiers, author results, and (16 centroids) the seven-ranking results.
SciNCL	History and candidate vectors; 1 or 16 centroids	Dense similarity. Bibliography tiers and author results.
Linear-Proximity	SPECTER2-Proximity vectors after learned linear maps	Supervised; see Table 15. Bibliography tiers, author results, and the seven-ranking results.

C.1 Linear-Proximity

The maps are initialized at the identity. Training scores candidates against individual history papers, whereas test-time ranking uses up to 16 centroids of the mapped history, as described above.

D Author-attestation protocol and retrieval results

This section gives the complete author-report evaluation, including how positive reports enter the analysis, the four target sets, and their retrieval results.

D.1 Collection, reports, and matching to the pool

Invitation text. The initial invitation read as follows, with identifying details removed for review; the shortlist lines held the numbered references.

> Dear Dr. [Last name],

Table 14: **Which ranking underlies each result.**

Ranking	Configuration	Results that use it
Base SPECTER2, 16 centroids	Base citation-trained encoder, without the proximity adapter	Matched-history analysis; same-case bibliography-versus-Context-Root comparison (Table 12).
SPECTER2-Proximity, 16 centroids (reference retriever)	Proximity-adapted encoder; fixed vectors, no learned maps	Main tables, bibliography coverage, author results, seven-ranking results.
Linear-Proximity, 16 centroids	Learned history and candidate maps over SPECTER2-Proximity vectors; mapped history clustered into up to 16 centroids	Main tables, author results, seven-ranking results (Table 15).
GritLM-7B, 16 centroids	GritLM-7B encoder; up to 16 centroids	Main tables, author results, seven-ranking results.
Qwen3-Embedding-8B, compact query	One vector from the compact query	Main tables and seven-ranking results.
SPLADE	Learned sparse matching on the compact query	Main tables and seven-ranking results.
BM25	Lexical matching on the compact query	Main tables and seven-ranking results.
Popularity	Citations received before P; no history	Main tables and seven-ranking results.
Hindsight test H / H+T / H+T+A	Reference retriever recomputed on one pool: history only; plus the next paper's title; plus a joint title-and-abstract vector in place of the title vector	Hindsight test only (Table 34).

Table 15: **Training and test-time scoring for Linear-Proximity.**

Component	Specification
Representation	Fixed SPECTER2-Proximity vectors for the researcher's history and for the candidate pool.
Learned scorer	Separate bias-free linear maps transform normalized history and candidate vectors; the transformed vectors are normalized. During training, the objective scores each candidate against every mapped history paper individually.
Training	Trained on training-split citations. The checkpoint is selected by nDCG@100 on CiteNew, averaged over validation cases; no test labels are used.
Selected checkpoint	Learning rate 3×10^{-5}, epoch 1.
Test-time scoring	Transform the history and candidate vectors, cluster the transformed history vectors into up to 16 spherical centroids, and score each candidate by its maximum cosine similarity to those centroids. Rank the candidates published before P; no second-stage reranker is used. Evaluation uses $K = 10, 100, 1000$.
Evaluation cases	All labeled test cases: CiteNext $N = 140{,}278$, CiteNew $N = 140{,}030$, and CiteNew-Isolated $N = 139{,}573$; a case without a ranking scores zero.

I'm a PhD student studying how prior work inspires new hypotheses and methods by tracing the intellectual lineage of impactful scientific papers. Specifically, we're building a curated dataset of "generative citations"—prior works that directly inspired a paper's central hypothesis or a key methodological move (beyond what citation counts can capture).

I read your paper, "[Paper title]," as part of this research, and I'm reaching out because authors are best positioned to say which citations were genuinely inspiring or formative.

Would you be willing to share which reference(s), if any, most directly inspired the core hypothesis or key methodological approach in your work? You can reply with: the number(s) (1–8) if listed below; 0 if none of these were primary inspirations; NA if you'd prefer not to respond; or ? if you're unsure.

Table 16: **The author-attestation campaign at a glance.** Stages are process steps, not nested counts (one response can yield several reports).

Stage	Count or description
Sources	Corresponding authors of papers in 25 journals, mostly life sciences (Table 17)
Invitation	Shortlist of up to 8 references from the surveyed paper's bibliography; unlisted works could be named
Reply options	Shortlist numbers, 0 (none of these), NA, ?, or any other work named by the author
Follow-up	Up to two reminders, seven days apart
Corresponding authors contacted	6,083 surveyed papers (5,460 unique emails)
Response rate	1,482 / 6,083 = 24.4% by surveyed paper
Responses captured	1,482
Confirmed reports after human review	2,677 (998 from the shortlist; 676 named outside it; 1,003 unclear)
Reports in evaluable cases	1,569 (1,438 with the work in the pool)
Evaluable cases	580 All, 460 Next, 359 New, 286 Isolated

Table 17: **Field composition of the 580 evaluable author-attested cases (Author-Root-All).**

Field family	Cases	Field family	Cases
Molecular and cell biology	217	Chemistry	17
Neuroscience	124	Psychology and cognitive science	17
Structural and chemical biology	58	Physics	10
Genetics and genomics	56	Materials science	10
Immunology	43	Computer science and AI	28
Total			580

Shortlist from your reference list:
1. [Reference] … 8. [Reference]

If the primary inspiration wasn't in this shortlist, even a URL/citation/name would be extremely helpful.

Your input will help improve scholarly recommendation systems so scientists can more easily find truly inspirational prior work.

[Name and affiliation removed for review]

Follow-ups repeated the shortlist and the same reply options. Table 16 summarizes the campaign.

Invitations asked corresponding authors which works shaped or inspired the surveyed next paper P^+, showed a numbered shortlist drawn from that paper's bibliography, and allowed other works to be named. The shortlist held at most eight references, selected from the full bibliography by a scoring rule that down-weights background and tool references and adds a small recency bonus; authors could reply with shortlist numbers, 0 (none of these), NA, or ?, and could name unlisted works by identifier, citation, or description. Non-responding authors received up to two follow-ups, seven days apart. Because the shortlist is pre-selected, attestations may favor listed works. Authors often looked beyond it: of the confirmed reports, 998 select shortlist items and 676 name works outside the shortlist, while for the remaining 1,003 the reply does not make clear which kind it is; 14 cases include an attested work that P^+ does not cite at all (Appendix D.3). The response rate was 24.4% by surveyed paper (Table 16). The 580 evaluable cases are a selective, author-collected evaluation set drawn from a different population than the citation-based benchmark, not a census of intellectual roots, and non-response may affect how representative the resulting reports are.

Panel composition. The 580 evaluable cases come from 25 journals, predominantly in the life sciences (Table 17), so author-attestation results describe these fields most directly.

The surveyed paper is P^+. For retrieval, P is the researcher's preceding paper and the query uses only history strictly before P. Neither P, P^+, their bibliographies, nor the response is a query

Table 18: **Author-collected cases by target availability and history.** A case is evaluated when at least one reported work is in the pool and the case has a usable history.

Target set	Cases with a report	At least one target in pool	All targets in pool	Evaluated (N)
Author-Root-All	628	596	521	580
Author-Root-Next	494	472	430	460
Author-Root-New	393	371	343	359
Author-Root-Isolated	316	298	280	286
Uncited available[†]	16	14	13	14

Before target availability is applied, 627 cases have a usable history for all eight retrieval conditions. [†]Uncited cases are a descriptive subset, not a fifth target set.

Table 19: **Confirmed reports and reports whose work is in the pool.** Counts are reports, not unique works or cases.

Target set	Confirmed reports	Reports with the work in the pool
Author-Root-All	1,569	1,438
Author-Root-Next	1,100	1,025
Author-Root-New	776	718
Author-Root-Isolated	580	537
Uncited available[†]	28	24

input. Unlike benchmark cases, these trajectories follow the researcher in any author position rather than as last author. Reports entered the analysis after each named work was identified and reviewed by a person; unresolved or non-scholarly reports were excluded. The 1,482 responses yield 2,677 confirmed reports of scholarly works, of which 1,569 fall in 628 cases with an Author-Root-All target. A missing response or an unselected shortlist item is treated as unknown, not as evidence against influence. The results describe responding authors who named at least one work, not how common intellectual influence is in general.

Each attested work is matched to a candidate in a fixed pool of 2,600,997 works, and counts as a target only if it was published before P. The pool adds the identified attested works and a year-stratified background sample to the benchmark pool, so these results are not directly comparable with bibliography-tier results. A case has a *usable history* when the researcher's papers before P include the text each retriever needs. A report whose work is not in the pool, or whose case lacks a usable history, is excluded rather than counted as a miss. The tables below count confirmed reports after this date check.

The four target sets are nested. *Author-Root-All* includes eligible works positively attested as shaping P^+. *Author-Root-Next* retains those same attested works only when cited in P^+. *Author-Root-New* removes works the researcher cited before P. *Author-Root-Isolated* further removes works cited before P by researchers who coauthored any of the researcher's papers in the preceding five years; those coauthors need not appear on P, and their citations may be older. Each result then uses the cases with a usable history and at least one target in the pool. Context-Root labels are a separate source of evidence and do not encode author intent.

The second table counts reports; the first counts cases.

D.2 Full author-panel retrieval matrix

The table reports Hit@100 for eight retrieval conditions across the four target sets. A hit means that at least one attested target appears in the first 100 candidates; each column has its own number of cases. With 286–580 cases per column, the table shows consistent patterns rather than a fine ranking of retrievers, and percentages in different columns are not on a common difficulty scale.

The prespecified 16-centroid-minus-1-centroid paired Hit@100 contrasts are calculated on identical cases within each target set. Table 22 gives the four Author-Root-All contrasts; the full set across

Table 20: **Reach on works new to the researcher under the two root evidence sources.** Both columns use the reference retriever (SPECTER2-Proximity, 16 history centroids).

Context-Root-New *Cited works new to the researcher whose citation context shows generative use*	**Author-Root-New** *Works new to the researcher that a responding author reported as shaping or inspiring the paper*
Test cases with a target in the pool and a usable history: $N = 6{,}704$ Benchmark candidate pool (2.33M)	Author-collected cases with a usable history and an attested work in the pool: $N = 359$ Benchmark pool plus attested works and extra background papers (2.60M)
Reach (Hit@100): 15.86% (1,063/6,704)	**Reach (Hit@100): 36.21%**

The retriever is the same in both columns; the evidence source, the cases, and the candidate pool differ, so the two percentages are not on a common scale of performance or difficulty.

Table 21: **Every representation loses Author-Root reach as attested works become newer to the researcher.** Hit@100 (%) for responding-author cases on the 2.60M-paper author pool; every column counts works that responding authors report as shaping their next paper. Tiers apply the same citation-history filters as the main results table. **Bold**: best per column.

Retriever	Representation	Author-Root-All *all attested* N=580	Author-Root-Next *cited in* P^+ 460	Author-Root-New *new to researcher* 359	Author-Root-Isolated *new to coauthors* 286
SPECTER2-Proximity	1 centroid	40.69	38.48	25.07	22.73
SPECTER2-Proximity	16 centroids (reference)	56.03	53.26	36.21	31.47
GritLM-7B	1 centroid	55.52	53.70	39.00	33.22
GritLM-7B	16 centroids	**59.14**	**57.61**	**41.78**	**34.27**
SciNCL	1 centroid	42.07	40.43	26.18	24.48
SciNCL	16 centroids	50.17	48.48	32.03	27.27
Linear-Proximity	1 centroid	38.28	36.74	22.01	18.88
Linear-Proximity	16 learned centroids	48.28	45.65	29.25	25.87

target sets is in Table 23. Intervals are percentile 95% intervals from 10,000 paired case-bootstrap replicates. They quantify paired sampling uncertainty for these panel cases and do not correct for response or reporting selection.

All four encoders gain from 16 centroids on Author-Root-All. On the newer tiers every gain stays positive, but the intervals for GritLM and SciNCL include zero on Author-Root-Isolated, and for GritLM on Author-Root-New.

D.3 Uncited author-attested cases: descriptive boundary

In 14 evaluated cases, at least one attested work is absent from the bibliography of P^+, while being a confirmed scholarly work in the pool, published before P, in a case with a usable history. Author reports can therefore reach beyond what a paper cites. We report only the count; identities and response text are withheld to protect participants, and the count is not a prevalence estimate.

Attested-work and citation-context overlap: not computable. We could not measure how often author-attested works are also Context-Root targets, because the two sets share no next papers: author-collected next papers are concentrated in 2024–2025, whereas Context-Root labels cover next papers concentrated in 2021–2024, and no shared identifier (DOI or PMID) links the two. No overlap estimate is reported.

D.4 Author-collected examples

Table 24 shows three author responses from the campaign in which the attested work was new to the researcher and to their recent coauthors, was in the benchmark pool, and was ranked using only the researcher's papers before P. Ranks come from base SPECTER2 with 16 history centroids

Table 22: **16-centroid-minus-1-centroid Hit@100 differences on Author-Root-All.** Differences are paired within the same cases and reported in percentage points.

Encoder family	Difference (pp)	Paired 95% CI (pp)
Proximity	+15.34	[+11.55, +19.14]
GritLM	+3.62	[+0.52, +6.72]
SciNCL	+8.10	[+4.31, +11.90]
Linear-Proximity	+10.00	[+6.38, +13.45]

Table 23: **Paired 16-centroid-minus-1-centroid Hit@100 differences across the four target sets.** Values are percentage points with percentile 95% case-bootstrap intervals (10,000 replicates); each pair uses the same target-specific cases as the matrix above.

Target set	Encoder family	Difference (pp)	Paired 95% CI (pp)
Author-Root-All	Proximity	+15.34	[+11.55, +19.14]
Author-Root-All	GritLM	+3.62	[+0.52, +6.72]
Author-Root-All	SciNCL	+8.10	[+4.31, +11.90]
Author-Root-All	Linear-Proximity	+10.00	[+6.38, +13.45]
Author-Root-Next	Proximity	+14.78	[+10.65, +18.92]
Author-Root-Next	GritLM	+3.91	[+0.43, +7.39]
Author-Root-Next	SciNCL	+8.04	[+3.70, +12.39]
Author-Root-Next	Linear-Proximity	+8.91	[+5.00, +12.83]
Author-Root-New	Proximity	+11.14	[+6.96, +15.60]
Author-Root-New	GritLM	+2.79	[−1.11, +6.69]
Author-Root-New	SciNCL	+5.85	[+1.11, +10.58]
Author-Root-New	Linear-Proximity	+7.24	[+3.34, +11.14]
Author-Root-Isolated	Proximity	+8.74	[+4.20, +13.29]
Author-Root-Isolated	GritLM	+1.05	[−2.80, +4.90]
Author-Root-Isolated	SciNCL	+2.80	[−2.10, +7.69]
Author-Root-Isolated	Linear-Proximity	+6.99	[+3.15, +11.19]

against the 2.33M-paper benchmark pool, not from the reference retriever on the author pool of Table 21. Descriptions are generic to protect participants. The examples illustrate how a history-based ranking can surface a work that later shaped a project, by transferring a method, broadening a theory, or expanding a shortlist; they are not a measured rate, and any claim about time saved would require a user study.

E MATCHED HISTORIES AND RETRIEVAL DEPTH

Does a researcher's own history contribute information beyond similar histories from other researchers? This section answers that question for the bibliography tiers, then gives the full depth metrics.

E.1 MATCHED HISTORIES OF OTHER RESEARCHERS

To test whether a researcher's history could be replaced by another researcher's, each case was paired with the histories of three other researchers. Each matched researcher had at least one paper before the case's cutoff, shared no earlier papers with the researcher, and fell in the same primary field (unknown field forming its own group), focal year, and history-size band, defined by $\lfloor \log_2(1+n_{\mathrm{pre}}) \rfloor$ for n_{pre} papers before P. In the content-matched version, each history's paper vectors were averaged and the three nearest eligible histories by cosine similarity were chosen, adding topical similarity to the field, year, and history-size match. The case's targets and candidate set were the same for the researcher's own ranking and the matched rankings.

Table 25 uses the test cases (about 28,000 per tier) with rankings for the researcher's own history and for all matched histories under both matching schemes. The metric is own-history nDCG@1000 minus the mean nDCG@1000 of the three matched histories, computed per case and averaged.

Table 24: **Three author-collected examples.** Attested works new to the researcher and their recent coauthors, with their rank under base SPECTER2 (16 centroids) on the 2.33M-paper pool.

Field	Next project	Attested work surfaced	Rank
Genomics	Local-ancestry association study in admixed populations	Prior local-ancestry method	14
Behavioral science	Complex-systems study of physical activity and behavior change	Prior behavior-change theory	54
Materials	Inverse design of hierarchical 3D architectures	Prior self-assembly pathway design	148

Table 25: **The advantage of a researcher's own history over three matched histories survives content matching (89–90% retained).** Own and matched columns give mean nDCG@1000 on the shared cases (base SPECTER2, 16 centroids); differences are paired means computed before rounding, so they can differ from the difference of the rounded levels by 0.0001. Retained advantage is the share of the advantage under field, year, and history-size matching that remains under content matching.

Tier	Shared N	Own nDCG	Matched nDCG	Own − matched	Retained advantage
CiteNext	28,446	0.0716	0.0076	0.0640	90.1%
CiteNew	28,393	0.0620	0.0072	0.0547	89.0%
CiteNew-Isolated	28,275	0.0599	0.0070	0.0529	89.0%

Positive values mean the matched histories do not fully substitute for the researcher's own. This evidence covers the bibliography tiers only; it does not identify a causal effect of author identity or extend to Context-Root or Author-Root targets. Confidence intervals are not shown for the displayed differences.

Content matching reduced the own-history advantage, relative to field, year, and history-size matching alone, by 0.0070, 0.0067, and 0.0065 nDCG points for CiteNext, CiteNew, and CiteNew-Isolated, respectively. The paired 95% intervals for these *changes* were [−0.00725, −0.00682], [−0.00695, −0.00653], and [−0.00673, −0.00631] (10,000 case resamples). These intervals are for the reductions, not for the own-minus-matched differences in the table.

E.2 Reach and coverage by depth

Here both metrics use represented cases: Hit@K (reach) follows Eq. (1) and macro Recall@K (coverage) follows Eq. (2), with E_i the targets of case i in its candidate set (published before P, with title and abstract) and I the represented cases. Cases without any target in the pool are excluded; 100 represented cases for which no ranking was produced are scored as misses.

Two conventions govern how a case with no eligible target in the pool is scored elsewhere in this paper. For the bibliography tiers, missingness is small (about 4% of CiteNew cases have no target in the pool), so scoring these cases as misses barely changes the reported reach and keeps the comparison realistic: any deployed system would also fail to find a work that is indexed nowhere. For Context-Root-New and the author-attested targets, missingness is much larger (about 40% of Context-Root-New cases), and counting them as misses would swamp genuine differences between retrievers with an artifact of data availability; those reach and coverage numbers are therefore restricted to represented cases only, as each table's caption states.

The next tables separate which cases are represented, which targets are in the pool, and retrieval performance. For CiteNew, 5,579 of 140,030 labeled cases have no target in the pool, leaving 134,451 represented cases. At $K = 100$, Hit@100 is 43,651/134,451 (32.47%), the share of cases reached; retrievable-target macro Recall@100 is 9.52%, the mean within-case fraction of eligible targets retrieved.

For completeness, Table 29 reports the same metrics with other denominators. All-case Hit includes labeled cases with no target in the pool, so it measures availability and ranking together; it matches Table 2. Observed macro coverage divides by all of a case's targets, including those not in the pool.

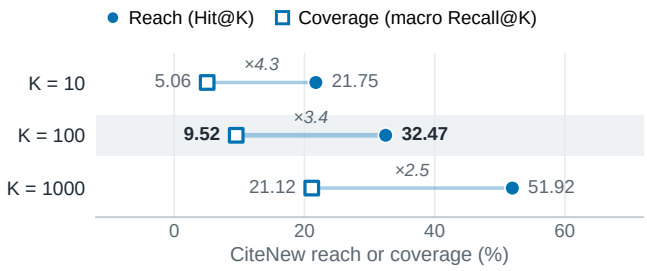

Figure 5: **Reach differs from coverage.** On CiteNew ($N = 134{,}451$ represented cases), the reference retriever's reach (Hit@K; cases with any target retrieved) and coverage (macro Recall@K; fraction of each case's targets retrieved) both rise from $K = 10$ to $K = 1000$ but remain distinct; gray $\times n$ gives reach divided by coverage. CiteNext and CiteNew-Isolated follow the same pattern (Appendix Table 28).

Table 26: **Represented cases by bibliography tier.** A case is represented when at least one target is in its candidate set, regardless of whether it is retrieved.

Tier	Labeled cases	No eligible target	Eligible cases
CiteNext	140,278	5,067	135,211
CiteNew	140,030	5,579	134,451
CiteNew-Isolated	139,573	6,200	133,373

Conditional coverage averages recall only over cases with at least one target retrieved. Micro recall pools all case–target pairs before dividing, so cases with more targets weigh more. None replaces the represented-case metrics above.

F Context-Root-New Retrieval and the Hindsight Test

This section reports the seven rankings on Context-Root-New, how many targets fall outside the pool, the hindsight test, and how reach varies with the researcher's history and with the targets.

F.1 Seven rankings and their pooled reach

Every ranking uses only candidates published strictly before the focal paper P. The main target is Context-Root-New: works cited by the next paper P^+ with a generative citation context and not cited by the researcher before P. Context-Root-Next keeps every generative-labeled work cited by P^+; Context-Root-Isolated also removes works cited before P by anyone who coauthored with the researcher in the five years before P (that citation may be older, and the coauthor need not appear on P). These two tiers serve as checks. A case is represented when at least one of its Context-Root-New targets is in its candidate set; 6,707 cases are represented. A miss means that no target appears in the first K candidates.

Table 31 reports all seven rankings on the same cases. Each uses only information available before P: the researcher's earlier papers or, for popularity, citation counts before P. The pooled result counts a case as reached if any of the seven reaches it; it summarizes separate rankings and is not a combined ranker. The hindsight test (Section F.3) is a separate procedure.

Formally, for represented case i and depth K, let $h_{ij}(K) = 1$ when ranking j places at least one Context-Root-New target in its first K results. Case i is reached by the pool when $u_i(K) = \mathbb{1}[\sum_{j=1}^{7} h_{ij}(K) > 0]$, and pooled reach is $\sum_{i \in I} u_i(K)/|I|$, where I is the set of represented cases ($|I| = 6{,}707$). A ranking that cannot be formed for a case has $h_{ij}(K) = 0$. At $K = 1000$, 3,619 / 6,707 = 53.96% of cases are reached by at least one ranking, and 3,088 / 6,707 = 46.04% are reached by none. This residual is specific to these seven rankings, this pool, and these cases. A further 4,482 of the 11,189 cases with a Context-Root-New target have none in the pool and are not counted. At $K = 100$, 1,923 / 6,707 = 28.67% are reached.

Table 27: **Target availability by bibliography tier.** A pair of case and target is in the pool when the target was published before P and has a title and abstract. Pairs not in the pool are not ranking misses.

Tier	Case–target pairs	In the pool	Lacking title or abstract
CiteNext	2,435,222	1,268,071	1,167,151
CiteNew	2,197,383	1,145,188	1,052,195
CiteNew-Isolated	1,912,598	1,014,921	897,677

Table 28: **Reach exceeds target coverage at every depth and tier.** Hit@K is the share of cases reached; macro Recall@K averages each case's fraction of retrievable targets retrieved.

Tier (N eligible cases)	K	Hit@K (reach)	Macro Recall@K (coverage)
CiteNext ($N = 135{,}211$)	10	27.45%	6.16%
CiteNext	100	37.40%	11.22%
CiteNext	1,000	54.73%	23.06%
CiteNew ($N = 134{,}451$)	10	21.75%	5.06%
CiteNew	100	32.47%	9.52%
CiteNew	1,000	51.92%	21.12%
CiteNew-Isolated ($N = 133{,}373$)	10	19.07%	4.96%
CiteNew-Isolated	100	29.31%	9.24%
CiteNew-Isolated	1,000	48.96%	20.46%

The SPECTER2-Proximity row is the reference retriever's ranking; the history-only condition of the hindsight test recomputes it, so its numbers differ slightly. BM25, SPLADE, and the compact-query Qwen3-Embedding-8B baseline share the compact query defined in Appendix C.

F.2 Targets outside the pool and dependence on individual rankings

Of the 11,189 test cases with a Context-Root-New target, 6,707 have at least one in the pool and 4,482 have none. Cases with none in the pool are not ranked; the 3,088 misses are represented cases with no target in any ranking's first 1,000 candidates. All-case Hit@1000 instead divides by every case with a Context-Root target.

Extending the ranked list from $K = 100$ to $K = 1000$ reaches 1,696 additional Context-Root-New cases (from 1,923 to 3,619), leaving 3,088 represented cases unreached. Dropping Linear-Proximity, the one supervised ranking, lowers pooled reach to 3,465 / 6,707 (51.66%) at $K = 1000$ and to 1,831 / 6,707 (27.30%) at $K = 100$, compared with 53.96% and 28.67% for all seven. Target availability and history availability are distinct: a target may be in the pool even when a retriever cannot form a query from the researcher's earlier papers.

F.3 Hindsight test: adding the next paper's title and abstract

The hindsight test holds the reference retriever's history representation and the candidate pool fixed. H uses history alone; $H + T$ adds the next paper's title to that history; $H + T + A$ keeps the same history and replaces the title vector with a joint encoding of the title and abstract. The added text exists only after the project, so the test measures dependence on it rather than offering a usable method, and it does not decompose the pooled misses above.

All three conditions are evaluated on the same cases with a usable history within each tier. The main comparison is Context-Root-New ($N = 6{,}704$); Context-Root-Next ($N = 7{,}848$) and Context-Root-Isolated ($N = 5{,}732$) are checks. Hit@1000 is the main depth and Hit@100 supports it. Table 34 reports all outcomes.

For Context-Root-New at $K = 1000$, adding the next paper's title changes Hit@1000 by +18.29 percentage points (95% CI [17.33, 19.26]); adding the abstract to the title changes it by a further +10.19 points (CI [9.34, 11.02]). From H to $H + T$, 1,273 cases go from miss to hit and 47 from hit to miss; from $H + T$ to $H + T + A$, 797 and 114. Intervals use 10,000 bootstrap replicates that

Table 29: **Reach and coverage with other denominators.** All-case Hit and observed coverage include targets not in the pool; conditional coverage includes only cases with a hit.

Tier	K	All-case Hit	Observed macro coverage	Conditional coverage
CiteNext	10	26.46%	3.01%	22.46%
CiteNext	100	36.05%	5.61%	29.99%
CiteNext	1,000	52.76%	11.78%	42.12%
CiteNew	10	20.89%	2.57%	23.25%
CiteNew	100	31.17%	4.90%	29.32%
CiteNew	1,000	49.85%	11.01%	40.67%
CiteNew-Isolated	10	18.23%	2.69%	26.00%
CiteNew-Isolated	100	28.01%	5.06%	31.53%
CiteNew-Isolated	1,000	46.79%	11.25%	41.78%

Table 30: **Secondary conditional observed coverage and target-level micro Recall.** Conditional coverage averages hit cases; micro Recall pools eligible target pairs and weights each target equally.

Tier	K	Conditional observed coverage	Eligible-target micro recall
CiteNext	10	11.39%	5.36%
CiteNext	100	15.56%	10.62%
CiteNext	1,000	22.33%	23.12%
CiteNew	10	12.31%	3.50%
CiteNew	100	15.73%	7.48%
CiteNew	1,000	22.08%	18.81%
CiteNew-Isolated	10	14.74%	3.27%
CiteNew-Isolated	100	18.06%	6.83%
CiteNew-Isolated	1,000	24.04%	17.21%

resample researchers. These gains describe only the hindsight test and cannot be read as recoveries of the pooled misses above.

The intervals reflect sampling of researchers only; they do not cover uncertainty from the model, the pool, the labels, or query construction, and they do not make the comparison causal.

F.4 Query construction and exclusions

For H, the query uses the researcher's papers with text published strictly before P. Their SPECTER2-Proximity vectors are grouped into up to 16 spherical centroids; a history of at most 16 papers uses each paper vector directly. $H + T$ adds one title vector for the next paper, and $H + T + A$ replaces it with a joint title-and-abstract vector. The history vectors and the pool are the same across conditions. A candidate's score is its highest similarity to any history or added vector. All conditions use the same cutoff and exclude P and P^+. Ties are broken by candidate identifier.

Three cases in each tier have too little history text to form a query and are omitted from all three conditions, leaving $N = 7{,}848$, 6,704, and 5,732 of the 7,851, 6,707, and 5,735 represented cases. All three omitted cases have a title and abstract for the next paper.

The main analysis keeps cases where the next paper's text may name a target directly. A check excludes the 258 Context-Root-New cases with an exact match on a target's title, author surname, corpus identifier, or DOI, or with very high title overlap; the title and title-plus-abstract gains become +18.27 and +10.11 points.

The history-only ranking is recomputed under the same procedure as the other two conditions and is therefore separate from the reference retriever's ranking in Table 31.

Table 31: **46.04% of represented Context-Root-New cases are missed by all seven rankings at $K = 1000$.** The 6,707 represented cases have at least one target in the pool. The six history-based rankings cannot be formed for three cases without a usable history; their denominator is 6,704. Popularity is available for all 6,707. Entries show Hit@K percentage with hit count in parentheses.

Ranking	N	Hit@100	Hit@1000
Popularity	6,707	2.28% (153)	6.62% (444)
BM25	6,704	13.01% (872)	23.97% (1,607)
SPLADE	6,704	13.87% (930)	24.16% (1,620)
Qwen3-Embedding-8B (compact query)	6,704	16.90% (1,133)	36.29% (2,433)
SPECTER2-Proximity (16 centroids)	6,704	15.86% (1,063)	29.46% (1,975)
GritLM-7B (16 centroids)	6,704	17.63% (1,182)	33.28% (2,231)
Linear-Proximity (16 centroids)	6,704	14.71% (986)	29.58% (1,983)
Pooled (any of the seven)	6,707	28.67% (1,923)	53.96% (3,619)

Table 32: **Inputs of the seven rankings.** Every input is available before the focal paper P; pooling does not combine rankings.

Ranking	Input and scoring
Popularity	Citations each candidate received before P; no history.
BM25	Compact query defined in Appendix C; lexical BM25 ranking over eligible candidates.
SPLADE	Same compact query; learned sparse expansion over eligible candidates.
Qwen3-Embedding-8B (compact query)	Same compact query encoded as one vector.
SPECTER2-Proximity (16 centroids)	Earlier-paper SPECTER2-Proximity vectors summarized by up to 16 spherical centroids; maximum similarity.
GritLM-7B (16 centroids)	Earlier-paper GritLM-7B embeddings summarized by up to 16 centroids.
Linear-Proximity	Learned linear maps over earlier-paper vectors, combined with up to 16 spherical centroids.

F.5 Reach by history size and recency

Pooled reach varies little across groups defined by the size and recency of the researcher's history. The outcome is whether any of the seven rankings reaches a Context-Root-New target by $K = 1000$, among the 6,707 represented cases; Context-Root-Next and Context-Root-Isolated are checks. History size counts the researcher's papers with text before P, capped at 20 (at most 4, 5–8, or at least 9). Recency is the time since the latest such paper: recent (≤ 0.167 years), middle (0.167–0.534 years), or less recent (> 0.534 years). These cutoffs are tertiles of the validation split, fixed before retrieval results were examined. The three cases without a recency value are shown separately.

Tables 35 and 36 report the share reached with 95% percentile intervals from 5,000 bootstrap replicates that resample researchers. Table 37 splits Context-Root-New cases into those reached by $K = 100$, first reached between ranks 101 and 1000, and not reached by $K = 1000$.

Reach by total number of earlier papers. The tables above use pooled reach and a capped count of papers with text. A complementary check uses the reference retriever alone and the uncapped number of the researcher's papers before P across all 6,707 represented cases, split into four bands (3; 4–5; 6–9; 10+ papers; N = 1,812, 1,751, 1,651, 1,493). CiteNew Hit@100 rises from 40.07% to 51.11% across these bands (an 11.04-point range), while Context-Root-New Hit@100 rises from 14.90% to 17.68% (a 2.78-point range)—comparable in size to the near-flat range already reported above. Both directions are confirmed by Spearman correlation across all 6,707 individual cases (CiteNew: $\rho = 0.083$, $p = 9.1 \times 10^{-12}$; Context-Root-New: $\rho = 0.026$, $p = 0.033$): reach rises with history size for both target types, but the association is markedly weaker for Context-Root-New than for CiteNew. This does not contradict the near-flat pooled figures above, which use a different

Table 33: **Cases with a Context-Root target, by availability and pooled retrieval.** Retrieval among available cases uses represented cases; all-case Hit@1000 uses every case with a Context-Root target.

Tier	Cases with a target	None in pool	Retrieved at $K = 1000$ among available	All-case Hit@1000
Context-Root-Next	12,786	4,935	4,777 (60.85%)	37.36%
Context-Root-New	11,189	4,482	3,619 (53.96%)	32.34%
Context-Root-Isolated	9,624	3,889	2,817 (49.12%)	29.27%

Table 34: **The next paper's text raises Context-Root reach in every tier and at both depths.** Each row uses the same cases, with a usable history, for all three conditions. Values are hits / N (Hit@K). The added title and abstract come from the next paper itself and exist only in hindsight.

Tier	K	History only	+ title	+ title and abstract
Context-Root-Next ($N = 7{,}848$)	100	1,793 / 7,848 (22.85%)	2,451 / 7,848 (31.23%)	3,199 / 7,848 (40.76%)
	1000	2,939 / 7,848 (37.45%)	4,174 / 7,848 (53.19%)	4,885 / 7,848 (62.25%)
Context-Root-New ($N = 6{,}704$)	100	1,059 / 6,704 (15.80%)	1,709 / 6,704 (25.49%)	2,411 / 6,704 (35.96%)
	1000	1,973 / 6,704 (29.43%)	3,199 / 6,704 (47.72%)	3,882 / 6,704 (57.91%)
Context-Root-Isolated ($N = 5{,}732$)	100	846 / 5,732 (14.76%)	1,427 / 5,732 (24.90%)	2,042 / 5,732 (35.62%)
	1000	1,581 / 5,732 (27.58%)	2,696 / 5,732 (47.03%)	3,295 / 5,732 (57.48%)

metric (pooled reach at depth 1,000) and different groups; taken together, they indicate that a larger history measurably helps recover ordinary future citations but only weakly moves whether a system recovers the generative root among them.

Within Context-Root-New, pooled reach ranges over 52.10–55.46% across history-size groups and 52.29–54.78% across recency groups, so these simple measures do not explain which cases are missed. This does not mean history is irrelevant: the comparison concerns only represented cases and these seven rankings.

F.6 Illustrative reached and missed cases

Table 38 shows seven test cases drawn from the reference retriever's ranking and the Context-Root-New targets, selected deterministically: among the 1,063 represented cases with a target in the reference retriever's top 100 (reached) and the 3,088 with no target in any of the seven rankings' top 1,000 (missed), cases were sorted by SHA-256 hash of their case identifier and the first three reached and first four missed were kept. Descriptions are generic to avoid identifying researchers; failure modes are our reading of each case, not a measured category.

The reached cases extend or apply a line of work already visible in the history, sometimes across a field boundary (tumor prognosis to a molecular cancer mechanism). All four missed cases show a topic shift between history and target, consistent with the finding that history size and recency do not explain which cases are missed (Appendix F.5): what the researcher does next is not always continuous with what they did before.

F.7 Target age and citation visibility

Complementing the history profiles, these observational summaries describe reach by the age of the targets and by whether they had been cited before P. Target age is the mean time from each available Context-Root target's publication to the date of P, averaged within each case. The age cutoffs (2.32

Table 35: **Reach at $K = 1000$ varies little with history quantity (52.10–55.46%).** Entries are cases reached by $K = 1000$ / group N (share, 95% CI resampling researchers).

Earlier-paper count	Tier	Reached / N	Share	95% CI
At most 4	Context-Root-New	1,503 / 2,885	52.10%	[50.28, 53.97]
5–8	Context-Root-New	1,138 / 2,052	55.46%	[53.31, 57.57]
At least 9	Context-Root-New	978 / 1,770	55.25%	[52.93, 57.55]
At most 4	Context-Root-Next	1,844 / 3,218	57.30%	[55.55, 58.99]
5–8	Context-Root-Next	1,505 / 2,412	62.40%	[60.41, 64.23]
At least 9	Context-Root-Next	1,428 / 2,221	64.30%	[62.30, 66.28]
At most 4	Context-Root-Isolated	1,224 / 2,560	47.81%	[45.88, 49.79]
5–8	Context-Root-Isolated	892 / 1,743	51.18%	[48.79, 53.49]
At least 9	Context-Root-Isolated	701 / 1,432	48.95%	[46.36, 51.47]

Table 36: **Reach at $K = 1000$ varies little with history recency (52.29–54.78%).** Entries are cases reached by $K = 1000$ / group N (share, 95% CI resampling researchers). Three cases without a usable recency value are shown separately.

Time since latest paper	Tier	Reached / N	Share	95% CI
Recent (≤ 0.167 years)	Context-Root-New	1,107 / 2,117	52.29%	[50.14, 54.38]
Middle (0.167–0.534 years)	Context-Root-New	1,217 / 2,223	54.75%	[52.66, 56.88]
Less recent (> 0.534 years)	Context-Root-New	1,295 / 2,364	54.78%	[52.75, 56.78]
No recency value	Context-Root-New	0 / 3	n/a	n/a
Recent (≤ 0.167 years)	Context-Root-Next	1,541 / 2,543	60.60%	[58.67, 62.50]
Middle (0.167–0.534 years)	Context-Root-Next	1,608 / 2,615	61.49%	[59.64, 63.30]
Less recent (> 0.534 years)	Context-Root-Next	1,628 / 2,690	60.52%	[58.68, 62.41]
Recent (≤ 0.167 years)	Context-Root-Isolated	829 / 1,773	46.76%	[44.46, 49.21]
Middle (0.167–0.534 years)	Context-Root-Isolated	941 / 1,882	50.00%	[47.73, 52.27]
Less recent (> 0.534 years)	Context-Root-Isolated	1,047 / 2,077	50.41%	[48.31, 52.56]

and 5.72 years) are Context-Root-New tertiles, fixed before outcomes were compared and applied unchanged to the other two tiers. A target is *uncited* if it had no citations strictly before P; a case is grouped as *uncited* if none of its available targets had been cited. Table 39 reports both measures for all three tiers.

For Context-Root-New, the raw difference between cases with at least one cited target and cases with none is -16.89 percentage points (95% CI [-20.14, -13.49], resampling researchers). A logistic regression of pooled reach at $K = 1000$ additionally controls for the number of targets and of available targets, mean target age, the log number of the researcher's earlier papers, availability of target abstracts, the log number of candidates, and availability of the next paper's text. In it, a one-standard-deviation increase in the share of uncited targets is associated with +0.363 log-odds of being reached (95% CI [0.300, 0.425]); the adjusted difference in predicted probability between a share of zero and a share of one is +22.81 points (CI [19.28, 26.34]). Field was not included because fewer than 90% of cases have a field tag. These associations are specific to these labels and this pool; they do not establish a cause or a mechanism.

A target-level replication confirms the same direction with a tighter estimate: of the 9,183 Context-Root-New target pairs in the pool, 12.78% (1,174/9,183) have zero observed citations before the focal cutoff, and the reference retriever's Hit@100 on these uncited targets exceeds the rate on cited targets by 29.06 percentage points (95% case-bootstrap CI [26.15, 32.02]). This holds among targets in the pool; having no citations is not the same as being hard for a dense retriever to find, since a target with no prior citations can still be textually close to the researcher's history. This target-level result and the case-level result above use different populations (target works in the pool versus represented cases) and different depths (K=100 versus K=1000); we report both rather than reconciling them into one figure. The advantage holds at K=1000 (29.85 points; CI [26.83, 32.79]) and for pooled reach rather than the reference retriever alone (48.98% versus 20.10% at K=100, +28.88 points, CI [25.88, 31.86]; 66.70% versus 46.37% at K=1000, +20.32 points, CI [17.41, 23.19]), so the effect is not specific to one encoder or ranking method.

Table 37: **Context-Root-New retrieval depth by history profile.** Shares of the 6,707 represented Context-Root-New cases reached by $K = 100$, first reached at ranks 101–1000, or not reached by $K = 1000$.

History profile	N	By 100	Ranks 101–1000	Not by 1000
At most 4 earlier papers	2,885	27.49%	24.61%	47.90%
5–8 earlier papers	2,052	30.36%	25.10%	44.54%
At least 9 earlier papers	1,770	28.64%	26.61%	44.75%
Recent history	2,117	27.35%	24.94%	47.71%
Middle recency	2,223	28.25%	26.50%	45.25%
Less recent history	2,364	30.29%	24.49%	45.22%
No recency value	3	—	—	—

Table 38: **Illustrative Context-Root-New cases, deterministically selected.** Reached: a target ranks within the reference retriever's top 100 (1,063/6,707 cases). Missed: no target appears in any of the seven rankings' top 1,000 (3,088/6,707). Cases are the first three reached and first four missed by SHA-256(case ID), not a representative sample.

Field	Researcher history → Context-Root-New target	Outcome
Reached by the reference retriever (1,063/6,707)		
Biology	Genomic biomarkers and vaccine-related effects → transcriptomic biomarkers applied to cancer	Rank 1
Medicine	Tumor location as a prognostic factor → gene-mediated mechanism in cancer progression	Rank 2
Physics	An extended field-theory model → related axion/dark-matter dynamics applied to cosmology	Rank 90
Missed by all seven rankings at $K = 1000$ (3,088/6,707)		
Tourism, management	Determinants of visitor counts → teacher competencies and performance through creativity	Topic shift
Marketing, consumer behavior	Continued consumption via planned-behavior constructs → entrepreneurship education and intention	Topic shift
Computer science	One-shot neural-network channel pruning → attention graph model for personalized recommendation	Topic shift
Computer science, vision	Causal deconfounding for image captioning → self-attention for point-cloud recognition	Topic shift

F.8 The bibliography versus Context-Root gap by field

Does the gap between bibliography and Context-Root reach (§4.2) hold within individual fields, or is it driven by one large field? Table 40 reports the paired CiteNew-versus-Context-Root-New Hit@100 difference on the same 6,707 represented cases and the reference retriever's ranking, broken out by each Semantic Scholar primary field with at least 20 cases (4,667 of 6,707 cases carry a field tag; the remaining 2,040 are untagged and excluded from this table only). Nine additional fields fall below the 20-case stability threshold (Political Science, 18; Materials Science, 16; Art, 15; Philosophy, 14; History, 10; Linguistics, 8; Law, 5; Geography, 4; Geology, 2) and are omitted.

The paired difference is positive in every field with $N \geq 20$: it ranges from +21.90 points (Education) to +42.77 points (Physics), and Computer Science, the largest field (2,341 cases), sits close to the middle of this range (+29.35 points) rather than driving it. The gap is therefore not an artifact of one dominant field, though its size varies roughly twofold across fields. Because a CiteNew case has many more targets than a Context-Root-New case (Table 11), part of each gap reflects target count; the declines from Next to New in the main text instead compare each evidence source with itself. We do not report field breakdowns of the author results or of the matched-history comparison, so this table establishes field generality only for the gap between bibliography and Context-Root reach, not for the personal-signal or project-text effects (§4.1, §4.3).

Table 39: **In every tier, cases whose Context-Root targets were uncited before P are reached more often.** Pooled reach at $K = 1000$ by target age and prior citation. Entries are cases reached / group N (share, 95% CI resampling researchers). Age cutoffs are the Context-Root-New tertiles, applied unchanged to the other two tiers.

Measure	Level	Tier	Reached / N	Share [95% CI]
Target age	Young	Context-Root-New	1,307 / 2,238	58.40% [56.38, 60.44]
	Middle	Context-Root-New	1,157 / 2,233	51.81% [49.71, 53.88]
	Old	Context-Root-New	1,155 / 2,236	51.65% [49.66, 53.72]
	Young	Context-Root-Next	1,350 / 2,246	60.11% [58.05, 62.21]
	Middle	Context-Root-Next	1,632 / 2,709	60.24% [58.38, 62.04]
	Old	Context-Root-Next	1,795 / 2,896	61.98% [60.19, 63.77]
	Young	Context-Root-Isolated	1,232 / 2,193	56.18% [54.07, 58.39]
	Middle	Context-Root-Isolated	837 / 1,881	44.50% [42.32, 46.75]
	Old	Context-Root-Isolated	748 / 1,661	45.03% [42.67, 47.41]
Citations before P				
	All targets uncited	Context-Root-New	558 / 811	68.80% [65.73, 71.94]
	At least one target cited	Context-Root-New	3,061 / 5,896	51.92% [50.63, 53.20]
	All targets uncited	Context-Root-Next	469 / 713	65.78% [62.32, 69.19]
	At least one target cited	Context-Root-Next	4,308 / 7,138	60.35% [59.20, 61.49]
	All targets uncited	Context-Root-Isolated	596 / 863	69.06% [66.01, 72.19]
	At least one target cited	Context-Root-Isolated	2,221 / 4,872	45.59% [44.21, 46.99]

Table 40: **The gap between bibliography and Context-Root reach is positive, with its 95% CI above zero, in every field with $N \geq 20$.** CiteNew and Context-Root-New Hit@100 (95% Wilson CI) and their paired difference (95% case-bootstrap CI, 10,000 resamples) on the same represented cases and the reference retriever's ranking, by Semantic Scholar primary field.

Field	N	CiteNew Hit@100	Context-Root-New Hit@100	Paired diff.
Physics	311	63.67% [58.18, 68.81]	20.90% [16.75, 25.76]	+42.77 pp [+37.30, +48.23]
Chemistry	57	50.88% [38.26, 63.38]	15.79% [8.54, 27.36]	+35.09 pp [+22.81, +47.37]
Biology	232	51.72% [45.32, 58.07]	16.38% [12.17, 21.68]	+35.34 pp [+29.31, +41.81]
Economics	75	50.67% [39.60, 61.67]	16.00% [9.40, 25.92]	+34.67 pp [+24.00, +45.33]
Sociology	38	50.00% [34.85, 65.15]	15.79% [7.44, 30.42]	+34.21 pp [+18.42, +50.00]
Agri. & Food Sci.	36	50.00% [34.47, 65.53]	19.44% [9.75, 35.03]	+30.56 pp [+16.67, +47.22]
Environmental Science	118	41.53% [33.04, 50.55]	11.02% [6.55, 17.94]	+30.51 pp [+22.88, +38.98]
Computer Science	2,341	43.83% [41.83, 45.85]	14.48% [13.11, 15.96]	+29.35 pp [+27.51, +31.23]
Medicine	371	41.78% [36.87, 46.86]	12.94% [9.90, 16.74]	+28.84 pp [+24.26, +33.42]
Engineering	219	47.03% [40.53, 53.64]	18.26% [13.71, 23.91]	+28.77 pp [+22.83, +35.16]
Mathematics	501	59.08% [54.72, 63.30]	32.53% [28.58, 36.76]	+26.55 pp [+22.75, +30.54]
Business	117	35.04% [27.00, 44.04]	9.40% [5.33, 16.05]	+25.64 pp [+17.95, +33.33]
Psychology	54	44.44% [32.00, 57.62]	20.37% [11.77, 32.90]	+24.07 pp [+12.96, +35.19]
Education	105	34.29% [25.91, 43.77]	12.38% [7.38, 20.04]	+21.90 pp [+14.29, +29.52]

G ENCODER EXPOSURE AND TRAINING-DATA OVERLAP

Publication-date masks restrict retrieval inputs and candidates to papers dated before each focal paper, but they cannot control what a pretrained encoder saw during training. We report two checks.

Training-free baseline. BM25 has no learned parameters. It reproduces the decline across citation-history tiers (32.71%, 28.68%, and 25.86% Hit@100 on CiteNext, CiteNew, and CiteNew-Isolated; Table 2) and reaches only 13.01% and 23.97% of represented Context-Root-New cases at $K = 100$ and $K = 1000$ (Table 31). The tier decline and the large share of unreached Context-Root-New cases therefore do not depend on encoder pretraining.

Training-data overlap for citation-pretrained encoders. We scanned the SciRepEval training triplets used by the SPECTER2 proximity adapter for exact matches to citation pairs in the test split (140,278 cases). Across the `cite_prediction` (676,150 triplets), `same_author` (67,493),

Table 41: **A pre-2021 encoder shows the same decline across tiers.** Hit@100 (%) with 16 history centroids on the earlier version of the benchmark; not comparable with the main tables.

Encoder	CiteNext	CiteNew	CiteNew-Isolated
SPECTER (2020)	39.52	30.15	25.72
SPECTER2	53.36	42.37	36.60

Table 42: **With target counts held equal, generative targets are reached more often than other new citations.** Case-level Hit@100 (%) on 5,702 cases of the earlier version, base SPECTER2 with 16 centroids. Matched rows report the mean $\pm$ standard deviation over repeated random draws (10 draws for the count-matched row).

Targets	Hit@100
Generative targets new to the researcher	20.5
CiteNew targets, matched on count per case	9.4 ± 0.3
CiteNew targets, matched on count, age, and prior citations	7.6 ± 0.4

and `high_influence_cite` (58,626) configurations, no triplet reproduces a CiteNext, CiteNew, or CiteNew-Isolated test pair. Context-Root-New targets are a subset of CiteNew targets, so no Context-Root-New pair overlaps either. The base SPECTER2 citation-prediction training set (6,197,963 triplets) was checked on an earlier version of the benchmark with a different, 167,568-case test split: 0.13% of CiteNew pairs overlapped, and removing them changed Hit@100 by at most about 0.0007. We have not repeated this check on the final split; base SPECTER2 is used only for the matched-history analysis and one same-case comparison. Qwen3-Embedding-8B and GritLM-7B do not document their training corpora at the level needed for an edge audit; for them, the training-free baseline above is the available check. On an earlier version of the benchmark, the original SPECTER encoder, trained before the test period, showed the same decline across tiers (Appendix H).

H Sensitivity Ablations

An encoder trained before the test period. Citation-trained encoders may have seen test-period citations (Appendix G). The original SPECTER encoder (Cohan et al., 2020) was released in 2020 and trained on earlier citation data, so it cannot have seen citations from the 2021–2023 test period. With 16 history centroids on the same pool, it reproduces the decline across tiers seen with SPECTER2 (Table 41): reach falls by 24% from CiteNext to CiteNew and by 35% to CiteNew-Isolated, against 21% and 31% for SPECTER2. The decline therefore does not depend on an encoder that could have seen test-period citations.

Generative targets with target counts held equal. A case with fewer targets is less likely to have one in the top 100, and cases have far fewer generative targets than CiteNew targets (Table 11). To separate the two, the earlier version compared, on the same 5,702 cases and the same ranking, the generative targets with CiteNew targets drawn to match each case's number of generative targets (Table 42). Generative targets were reached for 20.5% of cases at depth 100, against 9.4% for count-matched CiteNew targets and 7.6% when the drawn targets were also matched on age and prior citations. Counted per work, on the 5,652 cases with both kinds of target, the reach of generative targets was 17.6% at depth 100 and 24.3% at depth 1,000, against 4.7% and 11.3% for the other new citations of the same papers. Per work, then, generative targets are found more often than other citations; the lower case-level reach on generative targets reflects mainly that cases have fewer of them.

Number of history centroids. The main results fix the number of history centroids at up to 16 rather than tuning it. The value of CiteNext Hit@100 was 52.7% with 8 centroids, 53.4% with 16, and 53.6% with 64, so reach changes by less than a point above 8 centroids and 16 is not a tuned optimum.

Table 43: **Components of the release for this version and what each reproduces.**

Component	Released as	Reproduces
Case manifests (researcher, P, P^+, split, tier targets)	Identifier files with checksums	All populations and denominators
Candidate pool (2,330,779 papers) and per-case date masks	Identifier list and masking rule	Pool construction and eligibility
Context-Root labels and classifier adapter	Labels and LoRA adapter	Context-Root results; relabeling
Human-coded citation contexts (1,202; 300 double-coded)	Context identifiers and labels	Classifier validation
Author-attestation links (580 cases, all tiers)	Structured links only (Appendix I.3)	Author-Root results
Saved rankings (seven rankings, bibliography tiers, author results, matched histories)	Top-1,000 ranking files with checksums	Every retrieval metric
Construction, evaluation, and figure code	Versioned repository	End-to-end recomputation from identifiers
Paper text and metadata	Not redistributed; joined from an authorized S2ORC copy	Re-encoding and re-ranking

I Reproducibility, release, privacy, and licensing

MUSES is released as an identifier-only package: it reproduces every table on the final benchmark from saved artifacts without redistributing S2ORC text or raw author responses. This section describes what is public now, what the release for this version adds, how a reader reproduces the results, and how the author data is protected.

I.1 Release plan and reproduction path

Public now. Earlier anonymized v1.0.0 releases provide the benchmark identifiers and labels (`https://huggingface.co/datasets/anon-muses-neurips/muses`), companion citation-context labels (`https://huggingface.co/datasets/anon-muses-neurips/citeroots`), and a Qwen3-8B LoRA classifier adapter (`https://huggingface.co/anon-muses-neurips/citeroots-rhetoric-judge-qwen3-8b`). They were built from an earlier version of the benchmark, so their identifiers and labels differ from those evaluated here: they document the design but do not reproduce the reported numbers.

Release for this version. The release adds a dated, versioned package with per-file checksums, an inventory, and an asset-level license manifest (Table 43).

Two reproduction levels. From the release alone, a reader recomputes every metric on the final benchmark from the saved rankings and manifests (the checks in Appendix H used an earlier version and are not part of it); no result requires rerunning the GPT-5.4-mini teacher or repeating the author outreach, because the released Context-Root labels and author-attestation links are the evaluation inputs. Rebuilding the rankings additionally requires the S2ORC release named in Appendix A.1, joined by identifier, and the public models named in Appendix C. Where a run record lacks a detail such as the exact checkpoint or software revision, the release marks it unavailable rather than guessing it.

I.2 Compute and software provenance

The evaluated systems ran on heterogeneous hardware and software, so runtimes are not aggregated across runs. The release includes a run manifest recording, where available, the accelerator, framework and library versions, model, checkpoint, and tokenizer revisions, index settings, random seeds, input data version, ranking checksum, and evaluation settings.

I.3 Data access, privacy, and licensing

S2ORC papers, abstracts, citation contexts, and author metadata are source data and are not redistributed; the package documents how to join its identifiers to an independently obtained S2ORC copy.

Author responses were collected with consent for research use. The release includes only structured author-attestation data needed for evaluation: resolved paper identifiers, links from each surveyed paper to its attested works, tier assignments, and permissible structured metadata. It excludes raw email responses, verbatim response text, free-text rationales, correspondence, email addresses, and other identifying content. Before release, the author-derived files undergo a field-level privacy review covering schema minimization, consent scope, re-identification risk, and a removal path for authors who opt out. Nonresponse and unselected shortlist items are unknown, not negative attestation. Each released asset carries its own license entry, consistent with the terms of its upstream sources.