SELECTING PAPERS THE FUTURE WILL BUILD ON: UPTAKE-WEIGHTED COVERAGE OVER KNOWLEDGE THREADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting a compact set of papers to understand a research field is a set-selection problem: a good reading list should cover distinct lines of work while prioritizing knowledge likely to remain relevant to future research. Existing approaches based on citations, graph centrality, or LLM importance scores largely evaluate papers independently, making them prone to redundancy and retrospective bias. We introduce WCOV, an uptake-weighted coverage framework that decomposes papers into contribution-level claims and groups semantically related claims across papers into knowledge threads. We treat knowledge inheritance as the conceptual target and operationalize thread uptake as the number of later papers whose extracted claims semantically continue the thread. A citation-free thread-level predictor estimates uptake from semantic representations and thread metadata, after which weighted maximum coverage selects papers that cover high-uptake threads without double-counting shared content. The objective is monotone submodular, so greedy selection carries the standard $(1 - 1/e)$ approximation guarantee; an optional LLM-derived consensus prior trades future-uptake coverage against recovery of established landmark papers. We evaluate WCOV on graph neural networks and federated learning, selecting from papers through 2020, observing uptake in 2021–2022, and holding out 2023–2024 for evaluation. The uptake predictor achieves AUCs of 0.861 and 0.926, against 0.795 and 0.911 for extrapolating early uptake. At a budget of 50 papers, prediction-weighted selection covers 64.8% and 85.3% of held-out semantic-continuation events, against at most 56.3% and 65.4% for any external baseline. Adding the consensus prior improves landmark recovery at the corresponding coverage cost. These results establish future-uptake-weighted coverage as a principled alternative to independently ranking papers.

1 INTRODUCTION

A researcher entering an unfamiliar field, and increasingly an autonomous research agent surveying one (Lu et al., 2024; Boiko et al., 2023; Wang et al., 2024), must first decide what to read. Only a few dozen papers can be studied closely, so every later stage, from summarization to idea generation, inherits whatever that choice missed. A good reading list should do two things at once: recover the works a community regards as foundational, and cover the breadth of knowledge that subsequent research actually builds on.

Prevailing proxies are retrospective. Citation counts rank accumulated attention, which lags for recent work and favors surveys and datasets (Merton, 1968; Aksnes et al., 2019); PageRank inherits that bias and relies on sparse within-field citation graphs (Page et al., 1999; Chen et al., 2007); LLM judgments largely restate consensus in their training data (Zheng et al., 2023; Liang et al., 2024). Two structural problems follow. First, these signals track attention rather than knowledge flow: in our graph neural network (GNN) corpus, the rank correlation between citations at retrieval and citations from collected 2023–2024 papers is 0.336, and 77% of history papers receive none. Second, independent scores cannot discount overlapping content: among the top 20 federated learning (FL) papers, LLM ranking yields 18 shared-thread pairs versus 4 for our selection.

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.

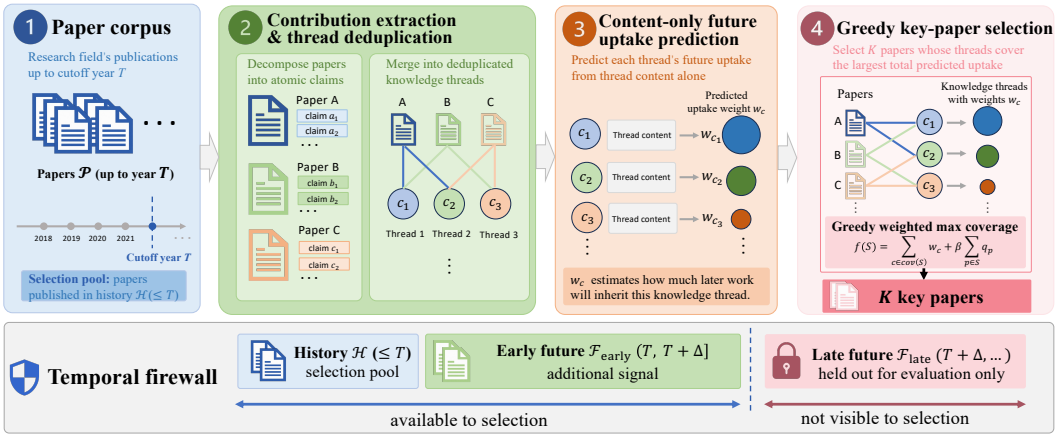

Figure 1: Overview. Papers are decomposed into deduplicated contribution-level knowledge threads, a bipartite structure in which one thread can span several papers. A content-based predictor assigns each thread a weight w_c, its predicted probability of future uptake, trained on the past and validated on the held-out future. Papers are then selected by greedy uptake-weighted maximum coverage.

We take a prospective view instead: a paper is key not because it is popular or highly rated today, but because it carries knowledge that later work inherits. We treat this inheritance as our conceptual target and make it measurable through thread-level semantic continuation: we work below the level of papers, decomposing each paper into contribution-level claims and merging semantically related claims across papers into knowledge threads, so that recurring content is represented once. The uptake of a thread is the number of later papers whose extracted claims restate, use, or extend it, our operational proxy for that inheritance. At selection time the future has not happened, so uptake must be predicted rather than measured, and our first empirical finding is that thread content alone suffices for that prediction.

Our framework, WCov, turns selection into weighted maximum coverage: it picks the K papers whose threads carry the largest total predicted uptake (Figure 1). The objective is monotone submodular, so greedy selection comes with the standard $(1-1/e)$ guarantee (Nemhauser et al., 1978), and covering the same thread twice earns nothing, which penalizes redundancy structurally rather than through a hand-tuned diversity term. An optional consensus prior, for instance an LLM importance score, enters the same objective through a single trade-off parameter β that traces a frontier between covering future-relevant knowledge and recovering expert-endorsed landmarks.

Evaluating a claim about future knowledge is delicate, because a method that optimizes an estimate of future uptake must not be scored on the very signal it optimizes. We therefore enforce a temporal firewall: the corpus is split at a cutoff year, the post-cutoff papers are split into an early window whose uptake is available to methods and a held-out late window, and the headline metric weights each thread by uptake measured strictly in that late window.

Our contributions are summarized as:

- We recast key-paper selection as a prospective question, which papers carry knowledge that later work inherits, operationalized as thread-level semantic continuation, and formalize it as uptake-weighted maximum coverage over deduplicated contribution-level knowledge threads: a monotone submodular objective that counts a shared thread once, carries a $(1-1/e)$ greedy guarantee, and admits an optional consensus prior through a single interpretable parameter β (Section 3).
- We show that a thread's future uptake is predictable from its content alone, at temporally validated AUC 0.861 and 0.926 across two domains against 0.795 and 0.911 for extrapolating past popularity, and we confirm this signal survives feature ablations and uptake labels rebuilt in an independent representation (Section 4.2).
- Under a temporal firewall that scores methods only on semantic-continuation events in a held-out later window, WCov covers 64.8% and 85.3% of those events at $K = 50$ against at most 56.3% and 65.4% for every external baseline in both domains, while β traces a

frontier that brings landmark recovery close to the strongest LLM baseline and keeps a significant coverage lead (Section 4.3).

2 Related work

Retrospective proxies and impact prediction. Bibliometrics ranks papers by citations and network centrality (Page et al., 1999; Chen et al., 2007), signals shaped by preferential attachment and recognition rather than quality (Merton, 1968; Aksnes et al., 2019). Future-impact prediction (Wang et al., 2013; Zhu et al., 2026; Ding et al., 2026) is prospective but still scores whole papers by expected attention; we instead predict contribution-level semantic uptake and select a non-redundant set.

Claim-level analysis and coverage selection. Our method combines two lines of work. Below the paper level, claim extraction and citation-intent classification isolate and label atomic assertions (Wadden et al., 2020; Jurgens et al., 2018; Cohan et al., 2019), Pramanick et al. (2026) annotate cross-paper claim relations and report that 63.5% of claims are never reused, and, closest to us, Jansen (2026) extract millions of contributions and predict prerequisite links under temporal backtesting. We share that granularity but predict whether a thread will be taken up at all, since exactly this sparsity is why uptake must be predicted rather than extrapolated. We then cast selection as submodular maximum coverage, whose greedy $(1 - 1/e)$ guarantee (Nemhauser et al., 1978; Minoux, 1978) underlies document summarization (Lin & Bilmes, 2011), result diversification (Carbonell & Goldstein, 1998), and benchmark selection (Smola, 2026), changing only what is covered, knowledge threads, and how it is weighted, by predicted future semantic uptake.

Representations, recommendation, and LLMs. Citation-informed encoders (Cohan et al., 2020; Ostendorff et al., 2022) and general sentence encoders (Reimers & Gurevych, 2019) supply the substrate for claim matching, which we treat as an implementation detail, and paper recommenders retrieve items for a query or profile whereas we cover a field. LLMs are increasingly used to judge (Zheng et al., 2023; Liang et al., 2024), survey (Wang et al., 2024), and drive autonomous research (Lu et al., 2024; Boiko et al., 2023); we treat an LLM as a strong baseline and an optional prior, combining its consensus signal with our predictor's signal about future semantic uptake.

3 Method

3.1 Problem setup

We are given a corpus partitioned by a cutoff year T into a history set $\mathcal{H}$ of papers published up to T, from which we select, and a future set $\mathcal{F}$ of papers published after T. We split $\mathcal{F}$ by time into an early window $\mathcal{F}_{\text{early}}$, whose uptake signal is available to methods, and a held-out late window $\mathcal{F}_{\text{late}}$, reserved exclusively for evaluation. Throughout we maintain a *temporal firewall*: every signal our method uses derives from $\mathcal{H} \cup \mathcal{F}_{\text{early}}$, while the headline metric is computed exclusively on $\mathcal{F}_{\text{late}}$, which no method observes. Given a budget K, the task is to choose $S \subseteq \mathcal{H}$ with $|S| \leq K$ that best covers contribution-level content with later semantic uptake. Figure 1 summarizes the pipeline.

3.2 Contribution-level knowledge threads

Papers are too coarse a unit: two may share a core contribution while differing in framing, and one may bundle several distinct contributions. We therefore extract, for each paper p, a set of atomic technical claims $A(p) = \{a_1, \ldots, a_{m_p}\}$ from its title and abstract with an LLM prompt that asks for self-contained, granular statements of what the paper contributes, yielding three to eight claims per paper (prompt in Appendix B). Each claim is embedded by a sentence encoder (Reimers & Gurevych, 2019) into a unit vector $\phi(a) \in \mathbb{R}^d$, and all subsequent matching operates on the cosine similarity $\text{sim}(a, b) = \phi(a)^\top \phi(b)$.

Claim typing. Rules assign claims to *method*, *theory*, *survey*, *other*, or *performance*. We exclude performance claims, such as "achieves state-of-the-art on citation benchmarks" (12% of GNN claims and 8% of FL claims; Table 2) because benchmark outcomes carry no transferable knowledge, reward verbose reporting under coverage, and are especially prone to false merges.

Anchor-gated deduplication. We then merge the remaining claims across papers into *contribution-level knowledge threads*, each a group of claims expressing the same underlying line of work. Embedding similarity alone does not suffice: if $\mathrm{sim}(a, b)$ and $\mathrm{sim}(b, c)$ both clear a threshold, naive transitive merging joins a and c however dissimilar they are, and repeated application chains topically adjacent claims into degenerate giant clusters. Two mechanisms prevent that.

(i) Anchor gate. Claims from a single field share a large generic vocabulary, so a high cosine similarity often reflects a shared topic rather than a shared technique. From each claim we therefore extract the tokens that name a specific method or resource, which we call its anchors $\alpha(a)$: hyphenated and CamelCase model names, all-caps acronyms, and parenthesized short names, upper-cased, minus a short stop list of field-generic terms such as GNN or FL that occur in hundreds of claims (patterns and stop list in Appendix B). We then define

$$\mathrm{compat}(a, b) \;=\; \neg\big[\alpha(a) \neq \emptyset \;\wedge\; \alpha(b) \neq \emptyset \;\wedge\; \alpha(a) \cap \alpha(b) = \emptyset\big], \tag{1}$$

so two claims are compatible unless both name specific artifacts and the named sets are disjoint. The gate is thus a veto layered on top of similarity rather than a merge criterion of its own, and it is inactive whenever either claim carries no anchor.

(ii) Complete-linkage clusters with a dual threshold. A cross-paper pair becomes a merge candidate when $\mathrm{sim} > \tau_{\text{cand}} = 0.70$ and the gate passes. Candidates are processed in decreasing similarity, and a claim joins a cluster only if its similarity to every member exceeds $\tau_{\text{strict}} = 0.65$ and the gate passes against every member. Demanding agreement with all members rather than any one eliminates transitive drift by construction. Appendix H reports threshold sensitivity.

The result is a bipartite structure between papers and threads. Write $\mathrm{threads}(p)$ for the threads a paper p contributes to and $\mathrm{papers}(c)$ for the papers containing thread c. A minority of threads are shared by two or more papers (Table 2), and the coverage objective treats repeated coverage of such a thread as redundant once any carrier is selected.

3.3 Definition and prediction of uptake

We operationalize *uptake* as automated semantic continuation between contribution statements extracted from titles and abstracts. Within this operationalization, a knowledge thread serves as a coverage unit for related contributions, while paper-level attribution and causal intellectual influence remain separate questions. Specifically, the uptake of a history thread c in a future window $\mathcal{W}$ is the number of distinct future papers with at least one claim matching the thread,

$$u_{\mathcal{W}}(c) \;=\; \big|\{\, q \in \mathcal{W} \;:\; \exists\, a \in c,\; b \in A(q),\; \mathrm{sim}(a, b) > \tau_{\text{cand}} \wedge \mathrm{compat}(a, b) \,\}\big|, \tag{2}$$

that is, the deduplication matcher applied across the time cut. We count papers rather than claims so that a single verbose follow-up cannot inflate a thread. Applied to the two future windows this yields an early-uptake count $u_{\text{early}}(c)$, which methods may use, and a late-uptake count $u_{\text{late}}(c)$, reserved for evaluation.

Content-based prediction. At selection time uptake must be predicted rather than observed, because a thread introduced just before T may attract little early uptake and still prove foundational later. We therefore learn $\hat{w}(c) \approx \Pr[u_{\text{early}}(c) > 0 \mid \mathrm{content}(c)]$ from content features alone: the mean of the member-claim embeddings, the thread type as a one-hot vector, and $\log(1 + |c|)$ for membership size. We binarize the label to any-uptake instead of regressing counts, because counts are heavy-tailed and their magnitude is partly a popularity artifact, so $\hat{w}(c)$ is a probability of any future semantic uptake rather than an expected number of matching papers: it ranks threads by whether later literature will continue them, not by how often. Positive rates are low, between 4 and 19% across domains and windows, so the model is class-weighted.

Two safeguards keep the predictor honest. The first is cross-fitting: scores come from 5-fold splitting, so every thread is scored by a model that never saw it, which keeps self-fitting bias out of the weights that drive selection. The second is deliberate simplicity: the predictor is an ℓ_2-regularized logistic regression with roughly 400 parameters, whose performance is stable under feature ablations and source-grouped folds in Section 4.2. The cross-fitted probability $\hat{w}(c) \in [0, 1]$ becomes the thread's selection weight.

As a purely observational alternative we also consider $w^{\text{obs}}(c)$, the early-window uptake count, which extrapolates recent momentum and which we call *momentum* throughout. It respects the

Algorithm 1 Uptake-weighted greedy selection

Input: history $\mathcal{H}$, threads with weights $\{w_c\}$, prior $\{q_p\}$, trade-off β, budget K
$S \leftarrow \emptyset$; $C \leftarrow \emptyset$ {C: threads covered so far}
while $|S| < K$ **do**
  $p^\star \leftarrow \arg\max_{p \in \mathcal{H} \setminus S} \left[\sum_{c \in \text{threads}(p) \setminus C} w_c + \beta\, q_p \right]$ {lazy evaluation with a max-heap}
  $S \leftarrow S \cup \{p^\star\}$; $C \leftarrow C \cup \text{threads}(p^\star)$
end while
return S (in selection order)

firewall as well and differs from $\hat{w}$ along one axis: it carries the uptake intensity that a probability weight abstracts away, but it scores every cold-start thread zero. We report both, and Section 4.3 separates their aggregate and cold-start behavior.

3.4 Selection as uptake-weighted maximum coverage

Given thread weights $\{w_c\}$ and a per-paper prior q_p (an optional consensus signal, e.g., a normalized LLM score), we select

$$\max_{S \subseteq \mathcal{H},\, |S| \le K} f(S), \qquad f(S) = \sum_{c \,\in\, \text{cov}(S)} w_c \;+\; \beta \sum_{p \in S} q_p, \tag{3}$$

where $\text{cov}(S) = \bigcup_{p \in S} \text{threads}(p)$ is the set of threads covered by S. The marginal gain of adding paper p to a partial selection S,

$$\Delta(p \mid S) \;=\; \sum_{c \,\in\, \text{threads}(p) \setminus \text{cov}(S)} w_c \;+\; \beta\, q_p, \tag{4}$$

has a direct reading: it is the total weight of the threads that this paper newly covers, plus a consensus bonus. A paper whose threads are already covered contributes only that bonus.

Proposition 1 *f in Eq. 3 is monotone and submodular for any nonnegative $\{w_c\}, \{q_p\}, \beta$; hence greedy selection returns S with $f(S) \ge (1 - 1/e)\, f(S^*)$.*

The coverage term is monotone submodular and the prior term is modular, which Appendix A verifies, and we optimize with lazy greedy (Minoux, 1978; Leskovec et al., 2007), exploiting the fact that submodularity keeps stale upper bounds on marginal gains valid.

The prior enters the objective, not the score. A tempting alternative is to blend a coverage score with a quality score after ranking, as a convex combination of normalized per-paper scores. That fails in a characteristic way: the quality term is available for every paper while coverage gains shrink as threads get covered, so the quality term dominates and the blend degenerates into quality ranking. Placing the prior inside the objective as a modular term keeps the trade-off within the greedy loop, where each step re-weighs the remaining coverage against the bonus. Normalizing q_p so that one unit of β equals the mean per-paper coverage gain $\sum_c w_c / |\mathcal{H}|$ then makes β dimensionless and comparable across domains.

The parameter $\beta \ge 0$ exposes the trade-off explicitly: at $\beta = 0$ selection is pure coverage of predicted future semantic uptake and runs fully offline with no LLM, and as $\beta \to \infty$ it collapses onto the consensus prior. Sweeping $\beta \in \{0, 0.5, 1, 2, 5, 10\}$ traces the frontier of Section 4.3. Setting $w_c = \hat{w}(c)$ gives our default WCov(pred), setting $w_c = w^{\text{obs}}(c)$ gives WCov(obs), and setting $w_c = 1$ for all threads gives an unweighted-coverage ablation that isolates the contribution of the uptake weights. Algorithm 1 summarizes the procedure.

4 Experiments

4.1 Setup

Domains and data. We evaluate on two fields with clear intellectual lineages and authoritative survey coverage: graph neural networks and federated learning. Both corpora are retrieved from

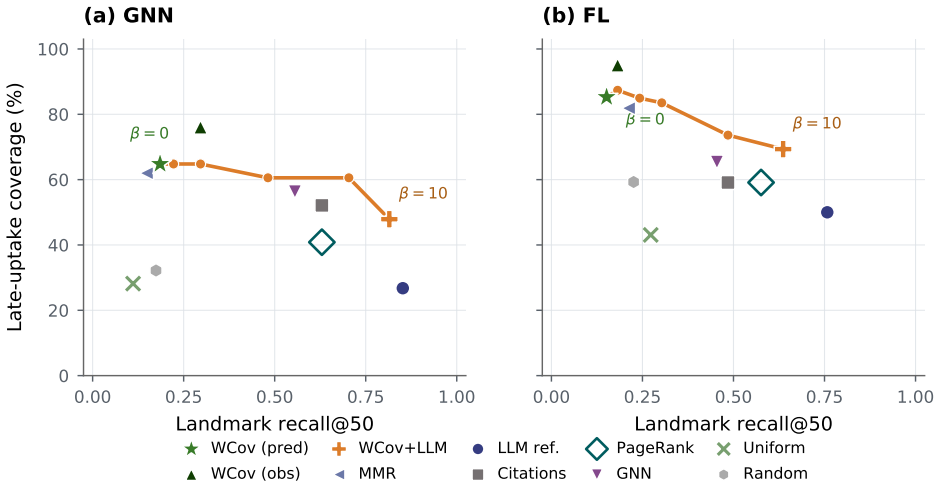

Figure 2: The coverage–consensus trade-off in both domains. The WCov+LLM β-frontier (orange) lies above every displayed baseline over the recovery range it spans: at any recovery level it reaches, it attains higher late-uptake coverage. Our offline WCov variants (stars/triangles) sit at the high-coverage end.

OpenAlex (Priem et al., 2022) by citation-ranked keyword search within publication year under fixed yearly quotas, over 2014–2024 for GNN and 2016–2024 for FL; in both, landmark papers missed by retrieval are added by title, so the whole landmark set is in the candidate pool by construction. We split each corpus at $T = 2020$ into history and future (2021–2024) and hold out 2023–2024 uptake for evaluation. Each domain yields a few hundred history papers and a few thousand claims, which deduplicate into a selection universe of 1,310 (GNN) and 857 (FL) threads carried by a history paper; Table 2 gives the full counts, and Appendix C details the identifier normalization that lets PageRank run on as complete a citation graph as possible.

Ground truth. To measure expert-consensus recovery we compile, per domain, landmark papers from three authoritative surveys and curated community lists (Appendix C), matched into the corpus by title: 27 (GNN) and 33 (FL) landmarks fall in the history window. This is the target of Recall@K. It records agreement with what the community already regards as foundational, which an LLM judge can restate directly, whereas our headline metric records thread-level semantic uptake observed in later work.

Metrics. *Expert-consensus recovery* is Recall@K against the landmarks, with 95% bootstrap confidence intervals over the landmark set (10,000 resamples). Our headline metric is *late-uptake coverage* (LUC), the share of late semantic-continuation events a selection touches,

$$\mathrm{LUC@}K \;=\; \frac{\sum_{c \in \mathrm{cov}(S_K)} u_{\mathrm{late}}(c)}{\sum_{c} u_{\mathrm{late}}(c)}, \tag{5}$$

reported in percent, where $u_{\mathrm{late}}(c)$ is thread c's uptake within the retrieved 2023–2024 literature, a quantity no method observes. It is count-weighted, so a thread inherited by many late papers contributes proportionally more than one inherited once, and its denominator is about $13\times$ smaller in GNN than in FL (Table 2), so GNN percentages move in steps of about 1.4 points and carry correspondingly wide intervals; we therefore report the two domains separately rather than pooling them. Performance threads are excluded consistently from both numerator and denominator. Alongside LUC we report a confound audit, the mean claims per selected paper, which exposes methods that game coverage by picking long abstracts, and pairwise thread overlap as a redundancy measure. Because LUC is weighted by our own matcher, we also validate it with a citation-derived metric that requires future work to both continue a thread and cite it, where our advantage persists (Appendix J).

Baselines. We compare citation count using each paper's full record at retrieval; PageRank over the normalized history-internal graph; per-paper LLM importance scoring; one-shot LLM set selection prompted for coverage and non-redundancy; a counterfactual GNN importance model over the paper–thread graph; and random selection.

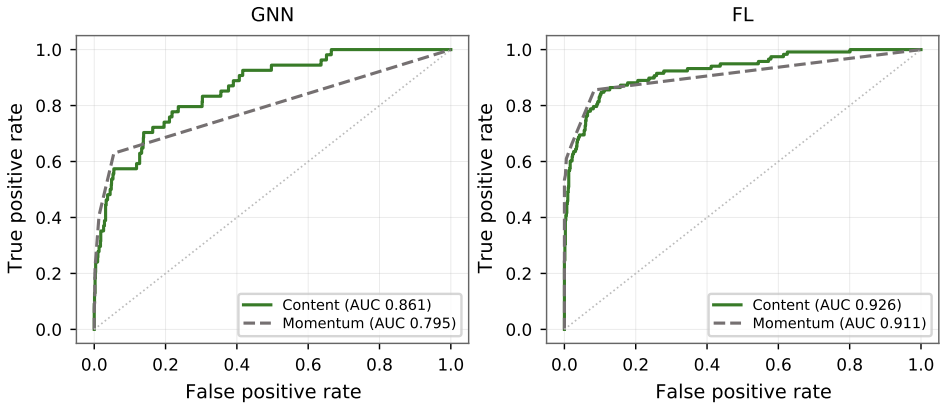

Figure 3: Predicting held-out 2023–2024 uptake in GNN (left) and FL (right). The cross-fitted content predictor (solid) reaches a higher AUC than the early-uptake momentum baseline (dashed) in both domains, with the larger margin in GNN where uptake is sparse.

Implementation. Claims are extracted and papers scored with DeepSeek-V4-Flash (temperature 0.1), claims are embedded with a 384-d MiniLM sentence encoder (Wang et al., 2020), and the uptake predictor is scikit-learn logistic regression. The pipeline is deterministic given a fixed seed, with all rankings tie-broken on paper identifier.

4.2 Future uptake is predictable from content

Our first finding is that whether a knowledge thread will be taken up by future work is predictable from its content alone. Against held-out 2023–2024 uptake the cross-fitted content predictor attains AUC 0.861 (GNN) and 0.926 (FL), where extrapolation from early-window uptake reaches 0.795 and 0.911; average precision is essentially tied in both domains (0.389 against 0.391, and 0.791 against 0.790), with every value far above its base rate (Appendix F). The AUC advantage is larger in GNN at a late positive rate of 4.1% and narrower in FL at 13.8%, so content adds the clearest discrimination where uptake is sparse (Figure 3).

Ablation. Under the identical cross-fitting protocol, full features (embedding, type and size) outperform either feature subset in both domains, and source-grouped folds preserve nearly the same AUC. The signal therefore survives both feature removal and a stricter separation of related threads across folds. The full ablation is in Appendix F.

Cross-representation check. The uptake labels come from similarity in the same embedding space that supplies the predictor's features, so we re-derive them from a representation sharing no parameters and no training corpus with it: TF-IDF cosine over the same claim strings. With the TF-IDF threshold set from the early-window positive rate, the same cross-fitted predictor leads the momentum baseline computed inside that independent representation by $+0.115$ (95% CI $[0.024, 0.204]$, $p = 0.012$) in GNN and by a smaller, non-significant $+0.034$ ($[-0.026, 0.090]$, $p = 0.240$) in FL. The partially overlapping label sets establish that the predictive signal is not confined to the shared embedding geometry (Appendix I).

4.3 Selection: coverage, consensus, and the trade-off frontier

Shown in Table 1, at $K = 50$, offline WCov(pred) covers 64.8% (GNN) and 85.3% (FL) of held-out semantic-continuation events, against at most 56.3% and 65.4% for any external baseline, and this ordering holds across the budget sweep (Figure 4). Paired bootstrap tests confirm the lead over PageRank and per-paper LLM scoring in GNN and over every external baseline in FL (Appendix E). Figure 4 also includes maximal marginal relevance (MMR; Carbonell & Goldstein, 1998) as an internal alternative selector: it uses the same predicted uptake weights as WCov(pred), adds a paper-similarity penalty, and selects its trade-off solely on early-window LUC@20. LLM set selection improves coverage over per-paper scoring by 18.3 and 5.8 points but trails WCov(pred) by 19.7 and 29.5 points; its selections also contain more overlapping paper pairs (39 versus 20, and 88 versus 63).

Table 1: Main results at $K = 50$, the budget at which every paired test of Appendix E is computed; per-domain tables with 95% bootstrap intervals on recovery are in Appendix D. R@50 is expert-consensus recovery, LUC@50 is held-out late-uptake coverage in percent, and cl/p is the mean claims per selected paper. Best entry per recovery and coverage column in **bold**.

	GNN			FL		
Method	R@50	LUC@50	cl/p	R@50	LUC@50	cl/p
Random (20 seeds)	0.174	32.2	6.12	0.226	59.3	5.29
Citation count	0.630	52.1	6.02	0.485	59.1	5.24
PageRank	0.630	40.8	5.88	0.576	59.1	5.10
LLM score	**0.852**	26.8	5.82	**0.758**	50.0	5.24
LLM set-select	0.815	45.1	5.90	0.697	55.8	5.04
GNN counterfactual	0.556	56.3	5.70	0.455	65.4	5.08
WCov (uniform)	0.111	28.2	7.10	0.273	43.1	6.34
WCov (obs uptake)	0.296	**76.1**	6.16	0.182	**95.1**	5.26
WCov (pred uptake)	0.185	64.8	6.54	0.152	85.3	5.96
WCov+LLM (β=10)	0.815	47.9	6.08	0.636	69.3	5.46

The gain comes from uptake weights, not the coverage framework. Coverage alone is not what drives the result: the unweighted ablation WCov(uniform) scores below chance in both domains and carries the highest claims-per-paper audit value in the table, since without weights the objective is maximized by long, claim-dense abstracts. Replacing uniform weights with predicted uptake raises coverage by 36.6 points in GNN and 42.2 in FL (Holm-adjusted $p = 0.010$ and $p < 0.001$) and returns the audit value toward the corpus mean, which attributes our gains to the uptake weights rather than to a generic diversity effect.

Attention and semantic uptake separate at selection time. Because LUC is count-weighted and the most frequently continued threads are carried by many papers at once, an unweighted random draw already reaches 32.2% (GNN) and 59.3% (FL) at $K = 50$. In FL, citation count and PageRank both reach 59.1%, essentially the chance level, while in GNN citation count rises to 52.1% but remains below WCov(pred) at 64.8%. The content-weighted selector clears chance by 32.6 and 26.0 points, respectively, showing that its gain is tied to future semantic continuation rather than retrospective consensus.

Prediction and observation play complementary roles. Because LUC is count-weighted while the predicted weight estimates only the probability of any uptake, observed early counts lead the content predictor by 11.3 points in GNN and 9.8 in FL. Both variants nevertheless lead every external baseline on coverage in both domains, while recovering fewer landmarks than the LLM and citation proxies do. Momentum exploits uptake intensity once it is visible; content prediction remains the only variant that can assign a nonzero score to cold-start threads, and in GNN it covers 42.9% of their late events against 28.6% for momentum.

The prior buys recovery with coverage along a smooth frontier. At $\beta = 0$ the objective contains no term that rewards consensus, so pure WCov(pred) optimizes coverage alone and ranks expert landmarks only incidentally: landmark lists and future semantic uptake are genuinely different targets rather than two views of one. Adding the LLM prior via β trades coverage for recovery along a frontier that lies above every displayed baseline over the recovery range it spans (Figure 2), with the LLM baselines sitting beyond the recovery range our β grid covers. By $\beta = 10$ recovery is statistically indistinguishable from the strongest LLM baseline (Holm-adjusted $p = 0.815$ and 0.578) while coverage stays 21.1 and 19.3 points above it ($p = 0.002$ and $p < 0.001$).

Case studies. The clearest added role of content is cold-start. In GNN, 1,207 threads have zero early uptake, yet 20 later activate and account for 21 late events; at $K = 50$, WCov(pred) covers 42.9% of those events against 28.6% for momentum. One selected example is a 2020 thread from *Design Space for Graph Neural Networks* about task-specific but transferable architectures: its source paper had 160 citations against a history median of 296, the content predictor assigned $\hat{w} = 0.505$, and a later GNN review continued the thread. In FL, a per-client homomorphic-encryption thread whose source paper the LLM rated 5/10 was matched by claims in 16 late papers ($\hat{w} = 0.999$), and a UAV-powered federated-learning thread rated 5/10 grew into a recognized subarea. The shared pattern

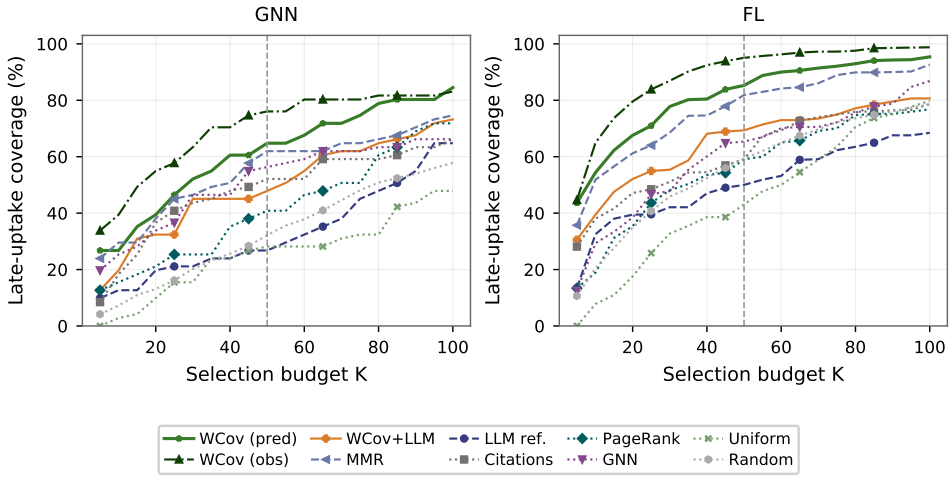

Figure 4: Late-uptake coverage as a function of the selection budget K. The unregularized uptake-weighted selectors achieve high late-uptake coverage across the evaluated budgets, while adding the LLM prior trades some coverage for landmark recovery. The vertical dashed line marks $K = 50$.

is that proxies grade a paper's present standing, while the content model can identify trajectories before uptake is observed.

Cross-domain replication. All findings above replicate across both domains, and FL is the more demanding test on every axis: a larger landmark set, a natively complete citation graph that makes the graph baselines maximally fair, and a far larger late-window evaluation sample. Threshold sensitivity results are in Appendix H.

4.4 Discussion

The citation graph is not the better substrate for predicting inheritance. Our counterfactual GNN importance model, trained over the paper–thread graph, matches a text-similarity control closely on early link prediction (AUC 0.815 against 0.819 in GNN and 0.716 against 0.693 in FL), yet trails the content selector on coverage (LUC@50 56.3% and 65.4% against 64.8% and 85.3%), even though it recovers more landmarks than pure coverage does and still fewer than the LLM baselines (Table 1). Its coverage rises with citation-graph density, from 2.9 edges per node in GNN to 6.9 in FL, yet even in the denser graph it stays below the content-coverage framework. We read this through the distinction of Section 1: a citation graph records attention flow, which aligns only weakly with what later work inherits, whereas thread content records the knowledge flow that selection for future inheritance depends on. Graph structure is therefore a complement that helps most where the graph is dense, not a substitute for content (Appendix G).

5 Conclusion

Key-paper selection is better posed prospectively than retrospectively: rather than ranking by accumulated attention, we ask which papers carry knowledge that later work inherits, which we operationalize as thread-level semantic continuation. Over knowledge threads that question is answerable from content, and uptake-weighted submodular coverage turns the predictions into a selection with a $(1 - 1/e)$ guarantee that covers more late-window semantic-uptake events than any external baseline in both domains, while a single parameter brings landmark recovery close to the strongest LLM baseline. A citation graph records attention flow; contribution content carries a predictable signal of future semantic uptake, a primitive that reading lists, survey generation, and research agents can build on.

Future work. Predicted uptake-weighted coverage could reward autonomous agents that produce reading lists, surveys, or research proposals by scoring contribution-level content likely to be continued. Its thread decomposition supplies per-step rewards, and the temporal firewall enables leakage-free validation. Extending the predictor to agent-generated contributions is the next step.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation, design or provide feedback on research methodology or experiments, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, clean and reformat dataset, or interpret results. Generating synthetic data sets and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated proof was verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work relies only on publicly available scholarly metadata from OpenAlex and involves no human subjects, personally identifying information, or sensitive data. Because our prospective signal is learned from a historical corpus, it can reflect that corpus's disciplinary and temporal biases, and it is intended to support literature triage, survey drafting, and research agents rather than to replace expert judgment or to act as a sole criterion for research evaluation or funding decisions. The contribution extraction and the LLM baselines use large language models, which can introduce errors or bias, and the temporal firewall keeps our evaluation leakage-free.

REFERENCES

Dag W. Aksnes, Liv Langfeldt, and Paul Wouters. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1):1–17, 2019. doi: 10.1177/2158244019829575.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Peng Chen, Huafeng Xie, Sergei Maslov, and Sidney Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007. doi: 10.1016/j.joi.2006.06.001.

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)*, 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)*, 2020.

Jianrong Ding, Jianyuan Zhong, Zhengyan Shi, and Qiang Xu. Fame: Forecasting academic impact via continuous-time manifold evolution, 2026. URL https://arxiv.org/abs/2605.07208.

Peter A. Jansen. The scientific contribution graph: Automated literature-based technological roadmapping at scale. *arXiv preprint arXiv:2605.15011*, 2026.

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.

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 420–429, 2007.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8), 2024. doi: 10.1056/AIoa2400196.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 510–520, 2011.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, pp. 234–243. Springer, 1978.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. ClaimFlow: Tracing the evolution of scientific claims in NLP. *arXiv preprint arXiv:2603.16073*, 2026.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3980–3990, 2019.

Alexander J. Smola. Submodular benchmark selection. *arXiv preprint arXiv:2605.02209*, 2026.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020.

Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. AutoSurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Hangxiao Zhu, Yuyu Zhang, Ping Nie, and Yu Zhang. Sciimpact: A multi-dimensional, multi-field benchmark for scientific impact prediction, 2026. URL https://arxiv.org/abs/2604.17141.

A Proof of Proposition 1

We show that $f(S) = \sum_{c \in \text{cov}(S)} w_c + \beta \sum_{p \in S} q_p$ is monotone and submodular whenever $\{w_c\}$, $\{q_p\}$, and β are nonnegative. Consider the coverage term $g(S) = \sum_{c \in \text{cov}(S)} w_c$ with $\text{cov}(S) = \bigcup_{p \in S} \text{threads}(p)$. For $S \subseteq S'$ and a paper $p \notin S'$, the threads of p left uncovered by S' are a subset of those left uncovered by S, so the marginal gains satisfy $g(p \mid S') \leq g(p \mid S)$, the defining diminishing-returns inequality of submodularity; since every $w_c \geq 0$, adding a paper never decreases g, giving monotonicity. The prior term $h(S) = \beta \sum_{p \in S} q_p$ is modular, and with $\beta, q_p \geq 0$ monotone nondecreasing, hence trivially submodular. Both properties are preserved under nonnegative addition, so $f = g + h$ is monotone submodular, and the $(1 - 1/e)$ guarantee for greedy maximization under a cardinality constraint follows from Nemhauser et al. (1978). □

B Prompts and extraction details

The domain string in each system prompt below is instantiated per field, as "Graph Neural Networks and deep learning on graphs" or "Federated Learning".

Claim extraction. Each paper's title and abstract are sent to DeepSeek-V4-Flash at temperature 0.1 with the system prompt below and a user message holding the title and the first $2{,}000$ characters of the abstract.

```
You are an expert in {domain}.
Extract the specific technical contributions (claims) from
this paper's title and abstract.

Each claim should be:
- A single, specific technical contribution or finding
- Self-contained (understandable without reading the paper)
- Granular: "introduces attention mechanism for neighbor
  aggregation in graphs" rather than "proposes a new GNN method"

Return ONLY a JSON array of 3-8 strings. Example:
["introduces spectral graph convolutions using Chebyshev
  polynomials",
 "achieves linear complexity in the number of graph edges",
 "demonstrates state-of-the-art on citation benchmarks"]
```

Claim typing. Types are assigned by rule patterns over the claim string (Section 3.2). Performance claims report benchmark outcomes rather than transferable knowledge; they are kept as singletons but excluded from clustering and from coverage accounting.

Anchor grammar and stop list. The anchor set $\alpha(a)$ of a claim (Section 3.2) is the union of four token classes, all upper-cased: hyphenated or compound acronyms matching `[A-Za-z]{1,4}[0-9]*-[A-Z][A-Za-z0-9]{1,8}` (e.g., ST-GCN, PB-GCN); CamelCase model names matching `[A-Z][a-z]+[A-Z][A-Za-z]+` (e.g., GraphSAGE, ChebNet); all-caps acronyms of length 3 to 10 (e.g., DCRNN, TGAT); and parenthesized short names of 2 to 15 characters. Tokens in a domain-generic stop list are dropped before matching, since they occur in hundreds of claims and carry no discriminative signal:

```
GNN, GNNS, GCN, GCNS, CNN, CNNS, RNN, LSTM, AI, ML, DL, NLP,
SOTA, GPU, API, II, III
```

extended per domain with a few field-generic terms: FL, SGD, DP, IID, NON-IID, and IOT in federated learning.

LLM importance baseline. The per-paper scoring baseline (Section 4.1) uses the system prompt below with the same title-and-abstract user message.

```
You are an expert in {domain}.
You will be given a paper's title and abstract. Rate how
```

```
important/foundational this paper is to the {domain} field
on a scale of 1-10.

Scoring criteria:
- 10: Groundbreaking foundational work that defined the field
      or a major subfield (e.g., the original GCN, GAT)
- 8-9: Highly influential work that introduced key concepts,
       widely-adopted methods, or theoretical insights
- 6-7: Solid contribution with notable impact, introduced
       useful techniques or benchmarks
- 4-5: Decent work with moderate contribution or incremental
       improvements
- 2-3: Minor contribution or small variation of existing work
- 1: Minimal significance to the field

Respond with ONLY a JSON object:
{"score": <number>, "reason": "<one sentence>"}
```

The LLM set-selection baseline (Section 4.1) replaces per-paper rating with a single prompt that presents all candidate titles and abstracts, and asks for a size-K subset maximizing coverage of distinct contributions while avoiding redundancy.

C Data collection and statistics

Both corpora are retrieved by citation rank within publication year (Section 4.1): we query OpenAlex (Priem et al., 2022) under fixed yearly quotas over 2014–2024 for GNN and 2016–2024 for FL. Both retain title, abstract, year, and reference list, and in both the landmark papers are fetched by title where search missed them; this adds two older GNN landmarks outside the bulk-retrieval years. We exclude one non-landmark GNN record whose returned title conflicts with its DOI metadata. The cutoff $T = 2020$ separates history from future, and the future set splits into an early window (2021–2022) and a held-out late window (2023–2024). Table 2 summarizes the result.

Table 2: Corpus statistics for both domains.

	GNN	**FL**
History papers (≤ 2020)	275	241
Future papers (> 2020)	203	321
Extracted claims	2,915	3,043
Performance claims (excluded)	351	242
Historical knowledge threads	1,310	857
Threads per history paper (mean)	5.26	4.64
Threads shared by ≥ 2 papers	7.7%	15.1%
Threads with late (2023–2024) uptake	54	118
late-uptake events (LUC denominator)	71	910
History-internal citation edges	786	1,670
Landmark papers in history	27	33

Citation-graph normalization. PageRank and the citation baselines run on the history-internal citation graph. Because references in OpenAlex mix identifier namespaces (OpenAlex IDs, DOIs, and titles), we resolve identifiers across namespaces before constructing the graph, yielding 786 GNN edges and 1,670 FL edges. This step strengthens the graph baselines and is independent of our content pipeline.

D Full results and coverage curves

Tables 3 and 4 repeat Table 1 one domain at a time, adding 95% bootstrap confidence intervals on recovery and keeping the chance floor as the only reference row. Figure 4 plots LUC@K over $K \in \{5, 10, 15, \ldots, 100\}$ for every method with a full ranking in both domains, on the same held-out late window as the headline metric and under the same firewall. Across the budget range, the

unregularized uptake-weighted selectors retain high coverage, while the hybrid exhibits the expected coverage–recovery trade-off, placing the $K = 50$ comparison within the same broader pattern.

Table 3: Main results, GNN domain, at $K = 50$, with 95% bootstrap confidence intervals on recovery. *Random* (20 seeds) is the chance floor and does not compete.

Method	R@50	R@50 95% CI	LUC@50 (%)	cl/p
Random (20 seeds)	0.174	n/a	32.2	6.12
Citation count	0.630	[0.444, 0.815]	52.1	6.02
PageRank	0.630	[0.444, 0.815]	40.8	5.88
LLM score	**0.852**	[0.704, 0.963]	26.8	5.82
LLM set-select	0.815	[0.667, 0.963]	45.1	5.90
GNN counterfactual	0.556	[0.370, 0.741]	56.3	5.70
WCov (uniform)	0.111	[0.000, 0.259]	28.2	7.10
WCov (obs uptake)	0.296	[0.148, 0.481]	**76.1**	6.16
WCov (pred uptake)	0.185	[0.037, 0.333]	64.8	6.54
WCov+LLM (β=10)	0.815	[0.667, 0.963]	47.9	6.08

Table 4: Main results, federated-learning domain, same protocol and budget (33 landmarks). The coverage ordering replicates, with WCov(obs) strongest here, where early uptake is dense.

Method	R@50	R@50 95% CI	LUC@50 (%)	cl/p
Random (20 seeds)	0.226	n/a	59.3	5.29
Citation count	0.485	[0.303, 0.667]	59.1	5.24
PageRank	0.576	[0.424, 0.758]	59.1	5.10
LLM score	**0.758**	[0.606, 0.880]	50.0	5.24
LLM set-select	0.697	[0.545, 0.848]	55.8	5.04
GNN counterfactual	0.455	[0.303, 0.636]	65.4	5.08
WCov (uniform)	0.273	[0.121, 0.424]	43.1	6.34
WCov (obs uptake)	0.182	[0.061, 0.333]	**95.1**	5.26
WCov (pred uptake)	0.152	[0.030, 0.273]	85.3	5.96
WCov+LLM (β=10)	0.636	[0.455, 0.788]	69.3	5.46

E SIGNIFICANCE TESTING

All tests use 10,000 paired bootstrap resamples at $K = 50$: for landmark recovery we resample the ground-truth set, for coverage the held-out late-window papers, recomputing every metric on each resample so that comparisons are paired, with Holm correction within each comparison family. Table 5 reports recovery tests for the β=10 hybrid. On coverage, WCov(pred) beats PageRank, the per-paper LLM, and uniform coverage in GNN ($p = 0.007$, $p < 0.001$, and $p = 0.010$), while its positive margins over citation count, set-selection, and the counterfactual GNN have wider intervals; in FL it beats all five external baselines and uniform coverage at $p < 0.001$. Relative to WCov(obs), its -11.3-point GNN difference is not significant ($p = 0.145$), while the -9.8-point FL difference is ($p < 0.001$).

Table 5: Paired bootstrap tests for landmark recovery at $K = 50$. The hybrid is statistically indistinguishable from the strongest LLM, citation, and PageRank baselines after Holm correction.

Comparison (R@50)	Domain	Δ	p
WCov+LLM vs LLM score	GNN	-0.037	0.815
WCov+LLM vs LLM score	FL	-0.121	0.578
WCov+LLM vs Citation	GNN	$+0.185$	0.603
WCov+LLM vs Citation	FL	$+0.152$	0.578
WCov+LLM vs PageRank	GNN	$+0.185$	0.603
WCov+LLM vs PageRank	FL	$+0.061$	0.632

F PREDICTOR ABLATION: FEATURES AND FOLD GROUPING

Table 6 gives the numbers behind Figure 5. All rows share the same logistic model and are evaluated by AUC and average precision against held-out 2023–2024 uptake. Full features (claim-embedding mean, thread type, and log cluster size) outperform either feature subset in both domains, while grouping threads connected through common source papers leaves the result nearly unchanged. The encoder is fixed to MiniLM throughout and is not itself ablated.

Table 6: Uptake-predictor ablation against held-out late uptake. Late positive rates are 4.1% (GNN) and 13.8% (FL). The full-feature logistic model is best or tied-best on AUC in both domains.

	GNN		FL	
Predictor	AUC	AP	AUC	AP
LR, full features (default)	**0.861**	0.389	**0.926**	0.791
LR, embeddings only	0.818	0.192	0.918	0.772
LR, metadata only (type+size)	0.767	0.340	0.747	0.484
LR, source-grouped folds	**0.861**	0.375	0.909	0.723
Momentum (early-count)	0.795	0.391	0.911	0.790

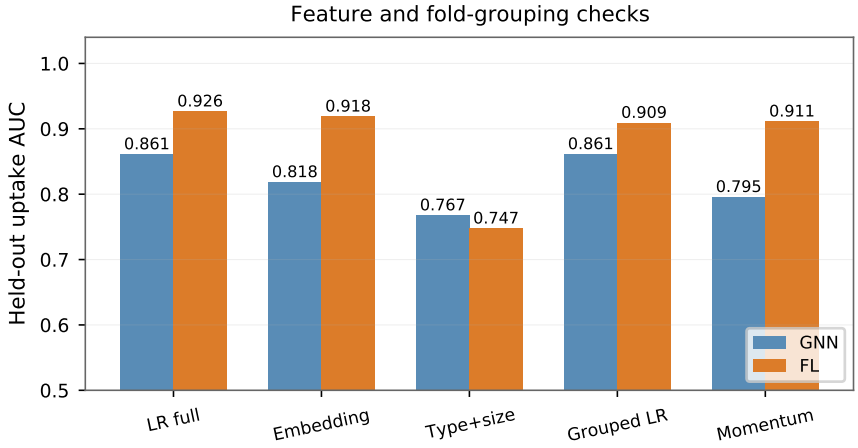

Figure 5: Uptake-predictor ablation (AUC vs. held-out 2023–2024 uptake, cross-fitted). Full features are best or tied-best in both domains, and source-grouped folds preserve the result.

G COUNTERFACTUAL GNN ANALYSIS

To test whether the citation graph is a better substrate than content (Section 4.4), we train a graph neural network on the paper–thread bipartite graph and use its learned paper importance as a selector, under the same firewall. Table 1 places it on both axes: below the content methods on coverage, and on recovery above pure coverage but below the LLM baselines. Head to head on recovery, the content selector trails it in GNN (paired $\Delta = -0.370$, $p = 0.006$) and FL ($\Delta = -0.303$, $p = 0.013$), confirming that graph structure is more aligned with landmark recovery. Its coverage rises with citation-graph density, which is higher in FL (6.9 against 2.9 edges per node), yet stays below the content-coverage framework in both domains; on early link prediction it remains close to a text-similarity control (AUC 0.815 versus 0.819 in GNN and 0.716 versus 0.693 in FL). Graph structure therefore encodes attention flow rather than the semantic-continuation signal used by our selector, consistent with the weak rank correlation between citation counts and later citations reported in Section 1.

H THRESHOLD SENSITIVITY

Table 7 sweeps both thresholds on the GNN corpus: the uptake match threshold τ_{cand} from 0.60 to 0.80, and the deduplication threshold τ_{dedup} that defines the threads themselves from 0.65 to 0.75.

The match threshold moves the LUC denominator from 558 late-uptake events at 0.60 to only 5 at 0.80, so absolute LUC values are comparable within a row but not down a column; the shared-thread count and largest cluster are fixed in the upper panel because clustering does not change. Across all eight settings, WCov(pred) leads the LLM baseline by 26.4 to 80.0 points. The content predictor also exceeds momentum at every setting except the loosest match threshold, where their AUCs are 0.863 and 0.865. The frozen $0.70/0.70$ configuration is not selected for the largest headline value: varying τ_{dedup} moves absolute coverage from 81.2% to 62.3%, while the advantage over the LLM baseline remains between 37.6 and 40.6 points.

Table 7: Threshold sensitivity on the GNN corpus. *Upper panel*: the uptake match threshold, with deduplication fixed at 0.70. *Lower panel*: the deduplication threshold, which defines the threads themselves, with the match threshold fixed at 0.70. The $0.70/0.70$ row is the frozen setting used everywhere else and appears in both panels. *mass* is the total late-uptake count, that is, the LUC denominator; the match threshold moves it by more than two orders of magnitude, so LUC (in percent) is comparable within a row and not down a column. AUC mom. is the observed early-count momentum baseline. *shared* counts threads carried by ≥ 2 papers and *maxC* is the largest such thread; both are properties of the deduplication threshold and therefore invariant in the upper panel by construction.

τ	AUC pred	AUC mom.	mass	LUC@50 pred	LUC@50 LLM	shared/maxC
Uptake match threshold τ_{cand}, with $\tau_{\text{dedup}} = 0.70$						
0.60	0.863	0.865	558	57.2	30.8	101/7
0.65	0.851	0.812	208	63.9	29.3	101/7
0.70	0.861	0.795	71	64.8	26.8	101/7
0.75	0.901	0.738	22	68.2	22.7	101/7
0.80	0.889	0.745	5	80.0	0.0	101/7
Deduplication threshold τ_{dedup}, with $\tau_{\text{cand}} = 0.70$						
0.65	0.901	0.777	64	81.2	40.6	144/11
0.70	0.861	0.795	71	64.8	26.8	101/7
0.75	0.863	0.820	77	62.3	24.7	63/4

I Cross-representation check: labels from an independent representation

Uptake labels are produced by MiniLM cosine similarity while the predictor's features are MiniLM embeddings, so we test whether predictability depends on that shared representation. We redefine uptake with TF-IDF cosine over the same claim strings, a bag-of-words representation with no learned parameters, no training corpus, and no geometry in common with MiniLM, holding everything else fixed: the anchor gate, the exclusion of performance-type claims, counting distinct future papers, the 5-fold cross-fitting protocol, and the MiniLM content features. The two cosine scales are not comparable, so we sweep the TF-IDF threshold τ_{tfidf} and report the entire sweep in Table 8. The headline row is the threshold whose early-window positive rate is closest to the frozen setting, which keeps label difficulty comparable: $\tau_{\text{tfidf}} = 0.35$ in both domains, at positive rate 0.069 against 0.079 in GNN and 0.166 against 0.190 in FL.

Three observations follow. First, the content predictor stays above chance on labels it did not help define, at headline AUC 0.786 and 0.689. Second, it exceeds momentum at six of nine thresholds in GNN and five of nine in FL, including both early-rate-matched operating points. Third, the two label sets overlap only partially, at Jaccard 0.184 (GNN) and 0.224 (FL) between TF-IDF and MiniLM late-uptake positives, with Matthews correlation 0.283 and 0.272, so prediction transfers across distinct label sets. A paired bootstrap over $1{,}000$ resamples of threads puts the content-minus-momentum AUC gap at $+0.115$ with 95% CI $[0.024, 0.204]$ ($p = 0.012$) in GNN and $+0.034$ with $[-0.026, 0.090]$ ($p = 0.240$) in FL. This experiment therefore establishes representation robustness, with a significant margin in GNN and a positive but non-significant margin in FL.

Table 8: Cross-representation sweep. Uptake labels are defined by TF-IDF cosine at threshold τ_{tfidf} while features remain the MiniLM content features. AUC pred is the cross-fitted content predictor against held-out late labels; AUC mom. is the observed early-count momentum baseline within the same TF-IDF representation; Jaccard is the overlap between TF-IDF and MiniLM late-uptake positives. Base-rate-matched headline rows in **bold**.

τ_{tfidf}	pos(early)	pos(late)	AUC pred	AUC mom.	Jaccard
GNN, MiniLM reference: pos(early) 0.079, pos(late) 0.041, AUC pred 0.861, AUC mom. 0.795					
0.20	0.546	0.331	0.709	0.719	0.107
0.25	0.279	0.167	0.732	0.689	0.152
0.30	0.144	0.079	0.762	0.711	0.179
0.35	**0.069**	0.037	**0.786**	0.671	0.184
0.40	0.036	0.016	0.797	0.704	0.119
0.45	0.019	0.009	0.871	0.745	0.100
0.50	0.011	0.003	0.919	0.873	0.055
0.55	0.006	0.002	0.974	0.999	0.056
0.60	0.005	0.001	0.962	1.000	0.019
FL, MiniLM reference: pos(early) 0.190, pos(late) 0.138, AUC pred 0.926, AUC mom. 0.911					
0.20	0.756	0.665	0.694	0.753	0.188
0.25	0.497	0.407	0.670	0.694	0.242
0.30	0.284	0.217	0.671	0.695	0.267
0.35	**0.166**	0.124	**0.689**	0.655	0.224
0.40	0.106	0.064	0.650	0.674	0.161
0.45	0.055	0.034	0.701	0.692	0.114
0.50	0.033	0.021	0.802	0.771	0.097
0.55	0.026	0.013	0.828	0.813	0.066
0.60	0.022	0.009	0.932	0.869	0.041

J External-metric validation: citation-derived coverage

Because LUC is weighted by our semantic matcher, we compare it with two citation-derived metrics based on real citation edges from held-out late papers.

Pure citation-attribution coverage (FCov). FCov credits a selection only when a late paper cites one of the selected papers, ignoring content entirely. On this metric we do not lead: at $K = 50$, WCov(pred) reaches FCov-hit 0.283 (GNN) and 0.583 (FL) against 0.509 and 0.750 for citation count (Holm-adjusted $p = 0.006$ and $p = 0.001$). The counterfactual GNN model, which also runs on the citation graph, reaches 0.528 and 0.725, slightly above citation count in GNN and slightly below it in FL. That is expected, since FCov rewards exactly the accumulated-citation signal citation count ranks by. The two notions of future relevance remain only partly aligned, at rank correlation 0.336 (GNN) and 0.609 (FL) between citation counts at retrieval and late citations, so paper attribution and thread-level semantic continuation measure different aspects of future relevance.

Intersection coverage (LUCv). LUCv counts a thread–late-paper event when the semantic matcher links the late paper to the thread and that paper cites at least one history paper carrying the thread. A selected set receives credit by covering that thread; the cited carrier need not be the selected paper itself. The denominator contains 20 events across 10 late GNN papers and 282 events across 59 late FL papers. At $K = 50$, WCov(obs) covers 90.0% (GNN) and 96.8% (FL) and WCov(pred) covers 90.0% and 91.5%, against 85.0% and 83.7% for citation count, 40.0% and 64.2% for the LLM, 80.0% and 68.4% for set-selection, and 20.0% and 60.3% for uniform coverage. In FL, where the citation-supported sample is larger, every gap is significant at $p < 0.001$: WCov(pred) leads citation count by 7.8 points, the LLM by 27.3, set-selection by 23.0, and uniform coverage by 31.2. The same point-estimate ordering holds in GNN, where the smaller sample produces wider intervals.