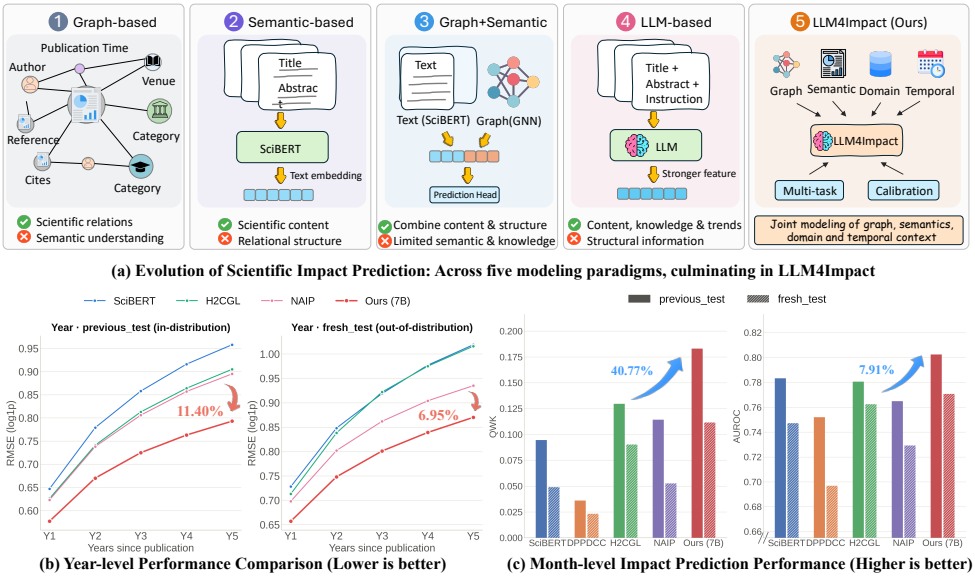
LLM4IMPACT: INTEGRATING HETEROGENEOUS INFORMATION FOR SCIENTIFIC IMPACT PREDICTION

Anonymous authors
Paper under double-blind review



(a) Evolution of Scientific Impact Prediction: Across five modeling paradigms, culminating in LLM4Impact

(b) Year-level Performance Comparison (Lower is better)

(c) Month-level Impact Prediction Performance (Higher is better)

Figure 1: **Overview of LLM4Impact and its performance.** (a) Scientific impact prediction across five paradigms: graph-based, semantic-based, graph+semantic, LLM-based, and LLM4Impact, which integrates graph, semantic, domain, and temporal information with LLMs; (b–c) Performance on both previous and fresh test sets demonstrates consistent gains across prediction settings.

ABSTRACT

Predicting the future impact of a newly published paper is challenging because it must be inferred from heterogeneous evidence available at publication time. Existing approaches often rely on a single source of information or combine multiple sources without accounting for their different predictive roles. In this paper, we present *LLM4Impact*, an evidence-aware method for scientific impact prediction that learns to represent, integrate, and calibrate heterogeneous information. LLM4Impact combines semantic, graph, LLM, and temporal representations, and injects graph information into a frozen LLM through continuous prefix tokens. A context aware gating mechanism adaptively weights different evidence, while a separate calibration module accounts for domain and temporal variation in citation scales. We further construct a large-scale benchmark dataset with 2 million papers, leakage-safe point-in-time heterogeneous ego graphs, temporal splits, and both year-level and month-level citation targets. Experiments show that LLM4Impact consistently outperforms strong semantic, graph, and LLM based baselines, with a 10.13% reduction in year RMSE on the in distribution test set and a 6.87% reduction under out-of-domain distribution. Our results reveal that the value of such evidence is context dependent: different papers benefit from different sources, while domain and publication time affect how evidence translates into citations. This finding motivates adaptive evidence selection and context-conditioned calibration rather than simply richer representations. We will release our code, benchmark, and an interactive website demonstration upon publication.

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.

1 INTRODUCTION

Predicting the future impact of scientific research is a fundamental problem in the science of science, with applications in identifying emerging research and supporting research evaluation (Bai et al., 2020; Xu et al., 2022; Xia et al., 2023; Zhang & Wu, 2024). A particularly challenging setting is cold start prediction, where a newly published paper has no citation history and its future impact must be inferred from information available at publication time. As shown in Figure 1(a), existing approaches have explored semantic content (Beltagy et al., 2019; Cohan et al., 2020), scholarly structure (He et al., 2023; Xue et al., 2024), temporal information (Wang et al., 2013; Jiang et al., 2021; Holm et al., 2022), and more recently LLMs (Zhao et al., 2025; de Winter, 2024). These signals are complementary, but their predictive value can vary across papers, domains, and publication periods. Effective cold start prediction therefore requires not only representing heterogeneous evidence, but also determining when and how each source should be used.

We identify three challenges that reveal why simply adding more information is insufficient. First, publication time can fall outside the range observed during training. A model based on discrete year embeddings cannot represent unseen publication years without learned embeddings, making temporal generalization difficult. Second, the predictive value of semantic, graph, and LLM evidence is instance dependent: sources that are informative for one paper may be less useful for another, making fixed or naive fusion suboptimal. Third, domain and publication time affect how evidence maps to citation outcomes, since citation scales vary systematically across contexts.

To address these challenges, we propose *LLM4Impact*, an evidence-aware method that jointly models semantic, heterogeneous graph, and LLM evidence under domain and temporal context for cold-start scientific impact prediction. These sources provide complementary signals: semantic representations capture paper content, heterogeneous graphs provide relational evidence, and LLM representations capture richer contextual signals. LLM4Impact preserves them as separate branches and injects graph information into the LLM through continuous prefix tokens, allowing structural and textual evidence to interact. Building on this multi-branch design, LLM4Impact introduces three components. First, a diagnostic-driven temporal encoder represents publication time with a continuous function and a zero-initialized residual, enabling generalization to unseen years. Second, a domain- and time-aware gated fusion module learns instance-specific weights over the semantic, graph, and LLM branches, reflecting their varying predictive value across contexts. Third, a decoupled context-conditioned calibration module adjusts predictions for domain and temporal variation in citation scales. Together, these components capture a central insight: effective impact prediction requires not only richer evidence, but also learning *which evidence to use* and *how to interpret it in context*.

For extensive training and evaluation, we construct a large-scale benchmark from the Semantic Scholar database containing 2 million query papers published between 2005 and 2020. For each query paper, we construct a leakage safe heterogeneous ego graph using only information available at the paper's publication time. The benchmark contains both an in distribution test set and a temporally shifted test set consisting of papers from later publication years. We evaluate both five year citation prediction and fine grained monthly prediction. This setting allows us to examine not only prediction accuracy, but also the ability to generalize when the publication time is outside the training range.

As shown in Figure 1(b-c), LLM4Impact consistently improves over strong semantic, graph based, and LLM based baselines. On the in distribution test set, five year RMSE drops from 0.790 to 0.710 (10.13%), and under temporal distribution shift, from 0.844 to 0.786 (6.87%). At the monthly level, LLM4Impact improves QWK by 40.77% in distribution and 24.44% under temporal distribution shift. These results show that the benefit of heterogeneous information persists across prediction horizons and temporal shifts. More importantly, ablations reveal that the LLM, graph, and calibration components contribute complementary gains, supporting the view that effective impact prediction depends on both evidence selection and context-specific calibration. We also provide an interactive demo for counterfactual probing, allowing controlled changes to venue, authors, publication time, or references while holding other information fixed to examine changes in predicted impact. In summary, our main contributions are as follows:

- **A new large-scale benchmark dataset for scientific impact prediction.** We construct and release a 2-million-paper benchmark with leakage-safe point-in-time heterogeneous ego graphs, temporal train/test splits, and support - for the first time - both year-level and month-level citation targets, together with data construction pipeline for future research.

- **An evidence-aware method for context-adaptive impact prediction.** We propose *LLM4Impact*, which dynamically selects and calibrates heterogeneous evidence under temporal and domain context, enabling both robust prediction under distribution shift and controlled probing of how contextual factors affect scientific impact.
- **Insights into heterogeneous evidence for impact prediction.** Across prediction granularities and temporal shifts, controlled ablations, pruning, and counterfactual probing reveal that evidence sources provide complementary and context-dependent signals, while the interactive demo enables controlled what-if analysis of how changes in authors, venues, references, and other contextual factors affect predicted impact.

2 Related Work

Predicting Scientific Impact. Predicting scientific impact has long relied on citation counts as a proxy for influence (Fu & Aliferis, 2008; Larivière & Gingras, 2010; Bai et al., 2019). Traditional bibliometrics like the h-index reflect established rather than future impact, prompting interest in article impact prediction based on early signals (Yang & Han, 2023; Vital Jr et al., 2025). Initial methods used handcrafted features, such as author reputation, venue rank, early citations, with regression or classification models, but struggled to capture the complexity of impact dynamics (Ibáñez et al., 2009). More recent approaches adopt data-driven and network-aware models, including temporal models that simulate citation growth via paper "fitness" and decay (Wang et al., 2013). Jiang et al. (2021) proposed HINTS, an end-to-end model predicting citation time series from publication time using pre-publication metadata and bibliographic networks, effectively addressing the cold-start problem and outperforming models dependent on years of citation data. Xue et al. (2024) introduced a GNN-based framework leveraging dynamic citation graphs and auxiliary tasks to yield interpretable and accurate predictions across paper lifespans and disciplines. These developments highlight the value of modeling scientific impact as a multifactorial process rather than a singular metric.

Heterogeneous Scholarly Networks Scholarly impact features are effectively modeled as heterogeneous information networks (HIN), where nodes (e.g., papers, authors, venues) and edges (e.g., citations, co-authorship) represent diverse entities and relations (Geng et al., 2022; He et al., 2023). Meta-path analysis enabled early relation-specific influence measures (Sun et al., 2011), followed by ranking methods that integrated content, venue, and publication networks (Tang et al., 2008). More recent work uses temporal GNNs and embedding alignment to track evolving influence and predict impact by embedding papers into historical network contexts (Holm et al., 2022). Advanced models apply relational GNNs with attention and temporal encoders to weigh author, venue, and content signals differently over time. HIN-based methods have also addressed collaboration prediction, topic emergence, and prestige estimation, confirming their strength for multi-scale scientific impact analysis (Zhao et al., 2025; de Winter, 2024; Arts et al., 2025; Jin et al., 2024). These models capture network structure and temporal context but underutilize the semantic richness of texts.

LLMs for Scientific Understanding. With the fast development of LLMs, researchers are increasingly exploring their ability to predict scientific impact from textual content alone (Cohan et al., 2020; de Winter, 2024; Zhao et al., 2025). Early efforts used simple representations like TF-IDF, but newer models leverage transformers trained on scientific corpora. SciBERT (Beltagy et al., 2019) improved classification and recommendation by capturing domain-specific language, while SPECTER (Cohan et al., 2020) showed that content-based embeddings correlate with scholarly relevance. Zhao et al. (2025) proposed a content-only LLM framework predicting impact from titles and abstracts in a double-blind fashion, achieving state-of-the-art performance using a field- and time-normalized metric, TNCSI_{SP}. Similarly, Vital Jr et al. (2025) showed GPT-based abstract embeddings could effectively identify highly cited papers, and even TF-IDF performed competitively, underscoring the role of topical relevance. de Winter (2024) found ChatGPT-4's qualitative scores on novelty, clarity, and engagement significantly correlated with later citations and Altmetric scores, suggesting LLMs can evaluate intangible manuscript qualities linked to impact. These studies show that LLMs excel at capturing intrinsic semantic merit and can complement or even rival traditional metadata-driven models. Combining LLM-derived content features with graph-based signals promises a holistic, multi-scale approach to modeling scientific influence.

3 DATASET CONSTRUCTION

To support systematic research on impact prediction, we construct a large-scale benchmark from the Semantic Scholar database (Kinney et al., 2023), along with an automatic pipeline for its construction.

3.1 CONSTRUCTION PIPELINE

Source corpus and query sampling. We build on the full S2AG snapshot, comprising 237,026,483 papers published between 1970 and 2025. Each paper is associated with static metadata, including publication information, venue, citation statistics, field-of-study tags, and abstract text. We define a *query paper* as a paper published between 2005 and 2020 with a non-empty reference list and an available abstract, ensuring both graph structure and textual content for downstream prediction. From the eligible pool, we uniformly sample 2,000,000 query papers without field-based stratification. We then split them temporally into a *historical* bucket (2005–2017) and a *fresh* bucket (2018–2020), with `train`, `val`, and `previous-test` drawn from the former and `fresh-test` from the latter. The temporal split is performed after the 2M papers are sampled, so the resulting historical and fresh subsets follow the natural temporal distribution of the sampled corpus. This temporal split enables evaluation of generalization to more recent papers and their citations.

Ego-graph construction and leakage-safe features. For each paper, we extract its 1-hop references and expand each 1-hop node with its top-25 references by citation count, forming a 2-hop ego-graph. To prevent temporal leakage, all citation counts are truncated to their values as of the query paper's publication year. Ego-graph statistics are reported in Table 1.

Table 1: **Statistics of our benchmark**: 2M papers with temporally separated test sets.

Split	Queries	Month Zero labels
Train	800K	722,510 (90.3%)
Validation	300K	270,810 (90.3%)
Previous Test	400K	361,196 (90.3%)
Fresh Test	500K	467,835 (93.6%)

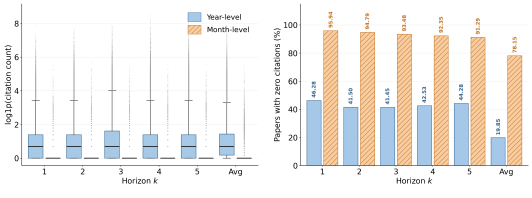
Query papers: 2M Years: 2005–2020
Ego-graph: 39.3M nodes, 1.18B edges
Metadata: author 99.5%, venue 46.1%, ROR 7.7%

Metadata cleaning and quality control. We augment each paper with author identities, canonical venues, and institutions, using S2AG's native venue links and ROR-based affiliation matching. We validate the resulting dataset through structural and referential consistency checks, including split disjointness, label monotonicity, citation-count validity, and key uniqueness. To rule out **memorization leakage**, all LLMs are queried without web search, and zero-shot LLMs perform worst among all methods (Tables 2–3), indicating that they cannot recall citation counts.

3.2 DATASET STATISTICS

Prediction targets. Following common formulations of citation-based impact prediction—single-horizon forecasting (Jiang et al., 2021), multi-step citation trajectories (Xue et al., 2024), and newborn-paper prediction (Zhao et al., 2025)—we consider both year-level ($k = 1, \ldots, 5$) and month-level ($k = 1, \ldots, 5$) citation counts as prediction targets. We provide both per-period and *cumulative* citation counts. For papers without a complete `publicationdate`, month-level targets are marked as `None` rather than zero, since the absence of a valid date does not imply zero citations.

Month-level targets are highly zero-inflated. Year- and month-level targets exhibit substantially different distributions (Figure 2). Unlike year-level counts, month-level targets remain zero for most papers at every horizon. Under such zero inflation, near-zero predictions achieve low regression error without discriminating among cited papers, so we complement regression metrics with threshold-sensitive classification metrics for the month-level task.



(a) Citation Distribution (b) Zeros citation ratio

Figure 2: **Statistics of citation numbers.**

4 METHOD

Scientific impact prediction is a cold-start problem: a newly published paper has no citation history and must rely on structural, semantic, and temporal evidence. Thus, we propose *LLM4Impact* (Figure 3), which integrates these signals through multi-branch and context-aware fusion.

Our design addresses **three distinct challenges** in cold-start impact prediction: representing unseen publication years, adapting to instance-dependent evidence reliability, and correcting domain- and time-dependent citation scales. We address them with **three corresponding components**: (i) a diagnostic-driven temporal encoder with continuous extrapolation and a zero-initialized residual, (ii) domain- and time-aware gated fusion, and (iii) a decoupled context-conditioned calibration module.

4.1 TASK FORMULATION

We formulate scientific impact prediction as a cold-start regression problem over a heterogeneous graph. Let $G = \langle N, E \rangle$ denote the heterogeneous citation graph, where nodes represent papers, authors, institutions, venues, categories, references, and publication time, and edges encode authorship, affiliation, venue, citation, and publication-time relations.

For a query paper p, all graph information is restricted to what was available at its publication time: its own citation count is zero, and neighboring papers are assigned only their point-in-time citation counts. This prevents future information from leaking into the prediction. Given p, the goal is to predict its cumulative citation counts over L future horizons:

$$Y_p = \{y_p^1, \ldots, y_p^L\}, \tag{1}$$

where y_p^ℓ denotes the cumulative citation count at horizon ℓ. We consider yearly prediction ($L = 5$) and monthly prediction ($L = 5$, month 1 through month 5). The model jointly leverages the query paper's heterogeneous graph, semantic content, and publication time.

4.2 MULTI-SOURCE REPRESENTATION LEARNING

Each query paper is represented by three complementary branches: semantic, graph, and temporal. Each branch produces a d-dimensional representation, with fusion deferred until Section 4.5.

Semantic branch. We encode the title and abstract of p using a frozen SciBERT (Beltagy et al., 2019) encoder followed by a two-layer projection head. The resulting representation is used as the semantic embedding of the query paper and also initializes paper/reference nodes in the graph branch, grounding semantic and structural representations in the same feature space.

Heterogeneous graph branch. We encode the query paper's 2-hop heterogeneous ego-graph using type-specific feature projections followed by a citation-aware GNN. Citation edges are weighted according to the sender's point-in-time citation count, while other relation types use mean aggregation. Type-aware attention pooling summarizes the heterogeneous neighborhood into a compact graph representation. During training, we additionally apply contrastive learning (van den Oord et al., 2019) between clean and popularity-augmented graph views.

4.3 GRAPH-TO-LLM PREFIX INJECTION

The graph representation provides structural evidence that is difficult to express through the LLM's text input alone. We therefore project the graph representation into the LLM hidden space and reshape it into L_p continuous prefix tokens (Li & Liang, 2021):

$$\mathbf{P} = \text{reshape}\left(\mathcal{P}_{\text{graph}}(\mathbf{e}_{\text{graph}})\right) \in \mathbb{R}^{L_p \times d_{\text{LLM}}}. \tag{2}$$

The prefix is prepended to the tokenized task instruction, title, and abstract directly in embedding space, while the LLM backbone remains frozen. The text positions are then pooled to obtain the LLM representation. Only the prefix projector, lightweight adaptation layers, and projection are trainable.

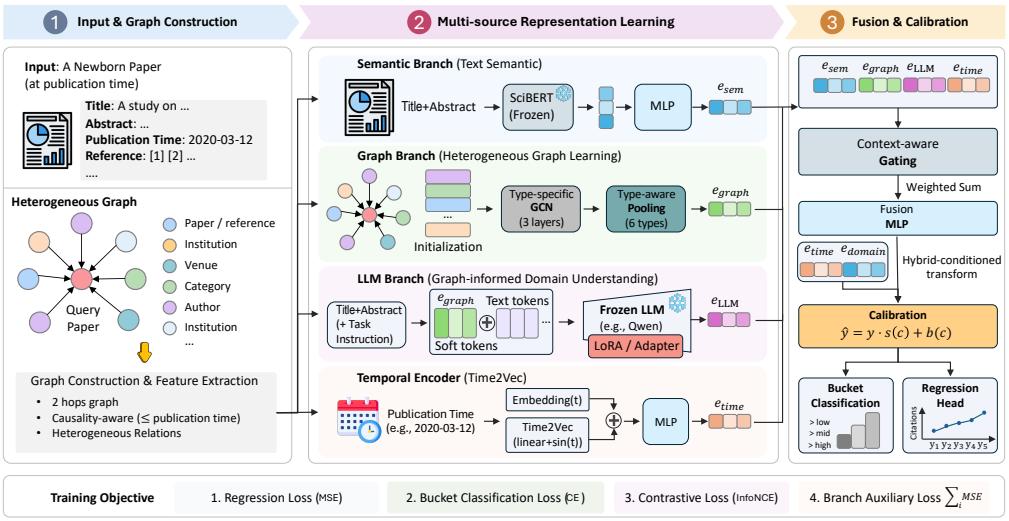


Figure 3: **Our proposed LLM4Impact.** A cold-start impact prediction method that integrates semantic, heterogeneous graph, and temporal evidence. Graph structure is kept as a separate branch and also injected into a frozen LLM via continuous prefix tokens, followed by domain- and time-aware gated fusion and decoupled context-conditioned calibration.

Prefix injection itself has been explored in other graph-centric tasks, but its effectiveness for citation-impact prediction has not been systematically established. The distinction is important because the query paper is cold-start and has no citation history of its own. We therefore compare embedding-space prefix injection against natural-language serialization of the same structural information. The prefix formulation consistently outperforms textual injection, supporting continuous structural conditioning as an effective interface for this setting.

4.4 Diagnosing and Fixing Temporal Encoding

Publication time is essential for impact prediction, but a standard lookup-based time embedding can fail under temporal generalization. Our fresh test covers papers from 2018–2020, while training uses 2005–2017. Thus, a year-based lookup assigns an untrained embedding to every fresh-test paper, since these years receive no gradient updates during training. We verify that this is the primary source of the fresh-test degradation rather than graph sparsity. Specifically, the error–graph-density relationship is nearly identical across the historical and fresh splits, ruling out graph density as the main differentiating factor and pointing instead to the unseen publication-time representation.

We therefore replace the lookup-only representation with a continuous, extrapolable time encoder (Kazemi et al., 2019) that applies uniformly to both year- and month-level prediction. Let τ_p denote the publication time of paper p, measured at the corresponding temporal granularity (year or month), and let t_p be its normalized value. We encode publication time as

$$\mathbf{e}_{\text{time}} = \text{MLP}_{\text{time}}\left([\sin(2\pi t_p \boldsymbol{\omega}); \cos(2\pi t_p \boldsymbol{\omega})]\right) + \text{Embed}_{,\text{time}}[\tau_p] \qquad t_p = \frac{\tau_p - \tau_{\min}}{\tau_{\max} - \tau_{\min}}. \tag{3}$$

This provides an extrapolable temporal representation, while the zero-initialized residual captures time-specific deviations for observed timestamps. Unseen timestamps simply fall back to the continuous base representation, enabling generalization across year- and month-level prediction.

4.5 Domain- and Time-Aware Gated Fusion

Evidence sources vary in their informativeness across papers. Thus, we learn an instance-specific gate conditioned on publication time and research domain. For paper p, we define $\mathbf{c} = [\mathbf{e}_{\text{time}} \| \mathbf{d}_p]$, where $\mathbf{d}_p$ is the mean-pooled category embedding. The gate produces normalized weights over branches:

$$\mathbf{g} = \text{softmax}\left(W_g[\mathbf{e}_1 \| \cdots \| \mathbf{e}_n \| \mathbf{c}]\right). \tag{4}$$

The gated branch mixture is then combined with the joint branch representation through a fusion MLP. Unlike fixed-weight fusion, the learned gate allows the model to adapt which evidence source to trust according to the paper's domain and publication era.

4.6 Decoupled Prediction Calibration

Evidence selection and prediction calibration address different problems. While fusion determines which evidence sources to trust, citation scales vary systematically across research domains and publication cohorts. We therefore apply a separate context-conditioned calibration module:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\text{fused}} \odot (1 + \text{MLP}_s(\mathbf{c})) + \text{MLP}_b(\mathbf{c}). \tag{5}$$

Zero-initializing the final calibration layers makes the module an identity transformation at initialization. Fusion therefore learns *which evidence to trust*, while calibration learns *how to rescale the estimate* based on domain and publication time. We train the model end-to-end with a composite objective:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda_{\text{aux}}\mathcal{L}_{\text{bucket}} + \lambda_{\text{cl}}\mathcal{L}_{\text{contrast}} + \lambda_{\text{branch}}\mathcal{L}_{\text{branch}}. \tag{6}$$

The main regression loss operates in $\log(1 + x)$ space. The auxiliary terms provide coarse citation-bucket supervision, graph-level contrastive learning, and branch-wise deep supervision, respectively. Detailed loss definitions and training settings are provided in Appendix B.

5 Experiment and Analysis

5.1 Experiment Setting

Baselines We compare against four representative baselines: SciBERT, a scientific text encoder; HINTS (Jiang et al., 2021), a dynamic heterogeneous network model; and NAIP (Zhao et al., 2025), an LLM-based impact prediction method. We additionally evaluate Qwen2.5-7B (Qwen et al., 2025), Qwen3-0.6B (Yang et al., 2025), Qwen3.5-4B, and Qwen3.5-9B (Team, 2026) via prompting and fine-tuning, excluding models larger than 9B due to dataset scale. Finally, we adapt H2CGL (He et al., 2023) and DPPDCC (Xue et al., 2024) to the newborn-paper setting to avoid citation leakage.

Evaluation Metrics We evaluate performance using regression, ranking, and zero-inflation-aware metrics. Year-level prediction is evaluated with RMSE. For month-level prediction, where 90.3% of `previous_test` labels are zero (see Table 1 and Section 3.2), we additionally report AUROC, PR-AUC, and QWK (Cohen, 1968), together with RMSE and NDCG@20. RMSE measures numerical prediction error, while NDCG@20 evaluates the quality of the top-20 ranked citation lists. Higher is better for AUROC, PR-AUC, QWK, and NDCG@20; lower is better for RMSE. Detailed definition of these metrics are in Appendix C.

5.2 Main Results

Tables 2 and 3 summarize the performance of our proposed LLM4Impact on the in-distribution and temporally shifted test sets, respectively. Overall, **LLM4Impact consistently outperforms existing approaches across both year- and month-level prediction, demonstrating the benefit of jointly modeling semantic, graph, and temporal evidence.**

On *previous_test*, our 7B model achieves the best performance at all five year-level horizons, reducing average RMSE from 0.790 of the strongest prior baseline, NAIP, to 0.710 (**10.13%** improvement). The gain grows with the prediction horizon, from 7.38% at Y_1 to 11.40% at Y_5, suggesting that structured and temporal evidence becomes increasingly valuable for longer-term impact prediction. The gains also extend to the highly zero-inflated month-level setting, where our model achieves the best trained-model performance across AUROC, PR-AUC, QWK, and RMSE, with a particularly large improvement in QWK (+40.77%). The smaller Qwen3-0.6B variant also outperforms prior baselines on most metrics, indicating that the gains are not solely attributable to the larger backbone.

Table 2: **Main results on previous-test (in-distribution).** Year-level results use per-horizon RMSE↓ in log1p space, while month-level results use zero-inflation-aware metrics due to 90.3% zero labels. Zero-shot LLMs are reported separately. Best results among trained models are shown in **bold**.

Method	Year (RMSE ↓)						Month				
	Y1	Y2	Y3	Y4	Y5	Avg	AUROC↑	PR-AUC↑	QWK↑	RMSE↓	NDCG↑
Traditional											
SciBERT (Beltagy et al., 2019)	0.647	0.779	0.858	0.916	0.958	0.839	0.783	0.406	0.095	0.296	0.562
HINTS (Jiang et al., 2021)	0.709	0.857	0.945	1.008	1.055	0.923	0.621	0.196	0.000	0.323	0.000
H2CGL (He et al., 2023)	0.626	0.741	0.813	0.864	0.905	0.796	0.781	0.417	0.130	0.296	0.246
DPPDCC (Xue et al., 2024)	0.665	0.791	0.866	0.920	0.958	0.846	0.752	0.364	0.036	0.305	0.383
LLMs-based											
NAIP (7B) (Zhao et al., 2025)	0.623	0.738	0.806	0.857	0.895	0.790	0.765	0.397	0.114	0.299	0.118
Qwen3-0.6B (zs) (Yang et al., 2025)	1.146	1.629	1.942	2.164	2.329	1.889	0.500	0.148	0.001	0.360	0.000
Qwen3.5-4B (zs) (Team, 2026)	1.189	1.494	1.694	1.820	1.912	1.642	0.648	0.249	0.099	1.245	0.090
Qwen3.5-9B (zs) (Team, 2026)	1.670	1.966	2.108	2.192	2.252	2.048	0.678	0.257	0.045	1.746	0.432
Ours (Qwen3-0.6B)	0.612	0.719	0.782	0.829	0.860	0.766	0.797	0.439	0.144	0.290	0.436
Ours(Qwen2.5-7B)	**0.577**	**0.670**	**0.725**	**0.763**	**0.793**	**0.710**	**0.802**	**0.450**	**0.183**	**0.286**	**0.566**
Improvement	*+7.38%*	*+9.21%*	*+10.05%*	*+10.97%*	*+11.40%*	*+10.13%*	*+2.43%*	*+7.91%*	*+40.77%*	*+3.38%*	*+0.71%*

Table 3: **Main results on fresh-test (out-of-distribution, temporally-shifted papers).** Same column design as Table 2. Best result per column among trained models in **bold**.

Model	Year (RMSE ↓)						Month				
	Y1	Y2	Y3	Y4	Y5	Avg	AUROC↑	PR-AUC↑	QWK↑	RMSE↓	NDCG↑
Traditional											
SciBERT (Beltagy et al., 2019)	0.728	0.848	0.919	0.977	1.019	0.904	0.747	0.399	0.049	0.355	0.588
HINTS (Jiang et al., 2021)	0.773	0.902	0.970	1.018	1.054	0.949	0.633	0.246	0.000	0.376	0.015
H2CGL (He et al., 2023)	0.713	0.838	0.922	0.975	1.016	0.899	0.762	0.422	0.090	0.351	0.259
DPPDCC (Xue et al., 2024)	0.743	0.869	0.940	0.989	1.037	0.921	0.697	0.346	0.023	0.367	0.229
LLMs-based											
NAIP (7B) (Zhao et al., 2025)	0.698	0.802	0.862	0.904	0.935	0.844	0.729	0.389	0.053	0.356	0.298
Qwen3-0.6B (zs) (Yang et al., 2025)	1.262	1.736	2.016	2.211	2.360	1.956	0.500	0.183	0.001	0.420	0.000
Qwen3.5-4B (zs) (Team, 2026)	1.268	1.591	1.794	1.916	2.001	1.734	0.651	0.298	0.112	1.313	0.039
Qwen3.5-9B (zs) (Team, 2026)	1.661	1.962	2.126	2.228	2.296	2.067	0.680	0.302	0.053	1.766	0.074
Ours (Qwen3-0.6B)	0.699	0.801	0.851	0.889	0.915	0.835	0.759	0.429	0.078	0.352	0.524
Ours (Qwen2.5-7B)	**0.657**	**0.748**	**0.801**	**0.839**	**0.870**	**0.786**	**0.771**	**0.441**	**0.112**	**0.347**	**0.657**
Improvement	*+5.87%*	*+6.73%*	*+7.08%*	*+7.19%*	*+6.95%*	*+6.87%*	*+1.18%*	*+4.50%*	*+24.44%*	*+1.14%*	*+11.73%*

More importantly, LLM4Impact maintains its advantage under temporal distribution shift. On *fresh_test*, where publication years are unseen during training, our 7B model remains the best trained approach at every year-level horizon, achieving an average RMSE of 0.786 and a 6.87% improvement over NAIP. It also achieves the best month-level performance across the reported metrics, including an 11.73% improvement in NDCG. This consistent performance on unseen publication years supports the effectiveness of our continuous temporal encoding and context-conditioned calibration. In contrast, zero-shot LLMs generally perform substantially worse on year-level prediction, highlighting that general-purpose language modeling alone is insufficient for cold-start scientific impact prediction.

5.3 Ablation Study

Table 4: **Ablation at year granularity.** Δ denotes the relative change compared with the full model.

Split	Model	Y1	Y2	Y3	Y4	Y5	Avg. RMSE	Δ
previous_test	**Ours (7B)**	**0.577**	**0.670**	**0.725**	**0.763**	**0.793**	**0.710**	–
	– Calibration	0.618	0.729	0.794	0.843	0.878	0.778	+9.6%
	– LLM branch	0.628	0.746	0.817	0.868	0.908	0.799	+12.6%
	– Graph branch	0.602	0.711	0.774	0.819	0.853	0.757	+6.6%
fresh_test	**Ours (7B)**	**0.657**	**0.748**	**0.801**	**0.839**	**0.870**	**0.786**	–
	– Calibration	0.702	0.814	0.881	0.937	0.976	0.868	+10.4%
	– LLM branch	0.708	0.829	0.905	0.962	1.005	0.888	+12.9%
	– Graph branch	0.679	0.777	0.835	0.878	0.908	0.819	+4.2%

Component Effectiveness. Table 4 evaluates the contribution of the major components on both in-distribution and temporally shifted data. Re-training method without LLM branch causes the largest degradation, increasing average RMSE by 12.6% on *previous_test* and 12.9% on *fresh_test*, demonstrating the value of LLM-based semantic evidence. Re-training method without calibration increases RMSE by 9.6% and 10.4%, respectively, supporting the benefit of decoupling prediction adjustment from evidence fusion. Re-training method without graph branch results in consistent increases of 6.6% and 4.2%, confirming the complementary role of structural evidence. The consistent effects across the two splits indicate that these components provide complementary benefits that persist under temporal distribution shift.

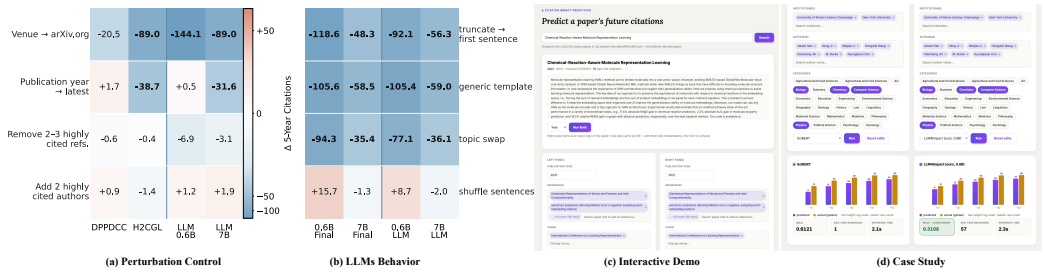


Figure 4: **Robustness, Interpretability, and Practical Validation of LLM4Impact.** (a) Contribution of different evidence sources; (b) LLM behavior under controlled perturbations; (c) Interactive exploration of citation predictions; (d) Case study comparing LLM4Impact with SciBERT.

Sensitivity to Perturbations. To examine what drives citation predictions and validate our multi-source framework, we conduct two complementary controlled perturbation studies on the same 20 computer science papers. As shown in Figure 4, (a) probes metadata sensitivity by perturbing authors, publication year, venue, and highly cited references while keeping other information fixed. Venue changes produce the largest effects across models, while author perturbations have much smaller impacts. The varying responses to publication year and citation structure further motivate domain- and time-aware gated fusion. (b) probes the LLM branch through sentence shuffling, truncation, generic replacement, and topic swapping. Content removal or replacement substantially changes both isolated LLM and final fused predictions, whereas sentence shuffling has a much smaller effect. These results indicate that the LLM branch captures content-level signals and provides complementary evidence to the graph and semantic branches, supporting our multi-source fusion design.

5.4 More Analysis and Discussion

Case Study: Interactive Sensitivity and Counterfactual Analysis We further provide an interactive demo for probing how contextual factors affect predicted scientific impact. Users can independently modify publication year, references, venue, authors, institutions, and research categories, while keeping the remaining information fixed, and compare the resulting citation trajectories across models. This enables controlled what-if analysis of how changes in individual factors alter the model's predicted impact, going beyond measuring sensitivity to input perturbations. The supplementary webpage provides additional screenshots and more case studies are in Appendix D.2.

Other Metrics. We further investigated the impact of alternative metrics on model evaluation. Following prior work, we extended our prediction target to Topic Normalized Citation Success Index(TNCSI) (Zhao et al., 2025), which normalizes citation counts relative to the citation distribution of publications in the same field, thus mitigating cross-disciplinary citation bias. Table 5 presents the RMSE scores via using both raw citation counts and the TNCSI metric, and their Spearman correlation coefficients (Spearman, 1961). We observe that despite differences in relative scores, the corresponding predictions exhibit a **strong positive correlation**. Besides, we find that the fresh test set continues to pose greater challenges compared to the previous one.

Table 5: **Comparison of RMSE and correlation *Cor.*** between C_log and C_TNCSI on *fresh* set.

Metric	Year_1	Year_2	Year_3	Year_4	Year_5	Avg
C_log	0.7170	0.7426	0.7892	0.8302	0.9514	0.7745
C_TNCSI	0.2316	0.2441	0.2556	0.2556	0.2532	0.2480
Cor.	**0.8004**	**0.8128**	**0.8095**	**0.8136**	**0.8128**	**0.8098**

6 Conclusion

In this paper, we presented *LLM4Impact*, a unified framework for cold start scientific impact prediction that integrates semantic, heterogeneous graph, LLM, and temporal evidence. By combining continuous temporal representation, context aware evidence gating, and context conditioned calibration, LLM4Impact consistently improves over strong baselines across yearly and monthly prediction tasks. Ablation and perturbation studies further show that different evidence sources provide complementary information and remain useful across changing publication contexts.

AI USE STATEMENT

Large language models are used in our early-stage writing for wording and grammar checking, as well as searching for missing literature. LLMs are not involved in the later iterations of the paper writing. Therefore, we do not consider LLMs as significant contributors to this paper.

REPRODUCIBILITY STATEMENT

We will publicly release our full code, including training, inference, and evaluation scripts, to ensure that all results in this paper are fully reproducible.

REFERENCES

Sam Arts, Nicola Melluso, and Reinhilde Veugelers. Beyond citations: Measuring novel scientific ideas and their impact in publication text. *Review of Economics and Statistics*, pp. 1–33, 2025.

Xiaomei Bai, Fuli Zhang, and Ivan Lee. Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407–418, 2019.

Xiaomei Bai, Fuli Zhang, Jin Ni, Lei Shi, and Ivan Lee. Measure the impact of institution and paper via institution-citation network. *IEEE Access*, 8:17548–17555, 2020.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3615–3620, 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 58th annual meeting of the association for computational linguistics*, pp. 2270–2282, 2020.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 10 1968. doi: 10.1037/h0026256.

Joost de Winter. Can chatgpt be used to predict citation counts, readership, and social media interaction? an exploration among 2222 scientific abstracts. *Scientometrics*, 129(4):2469–2487, 2024.

Lawrence D Fu and Constantin Aliferis. Models for predicting and explaining citation count of biomedical articles. In *AMIA Annual symposium proceedings*, volume 2008, pp. 222, 2008.

Hao Geng, Deqing Wang, Fuzhen Zhuang, Xuehua Ming, Chenguang Du, Ting Jiang, Haolong Guo, and Rui Liu. Modeling dynamic heterogeneous graph and node importance for future citation prediction. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 572–581, 2022.

Guoxiu He, Zhikai Xue, Zhuoren Jiang, Yangyang Kang, Star Zhao, and Wei Lu. H2cgl: Modeling dynamics of citation network for impact prediction. *Information Processing & Management*, 60 (6):103512, 2023.

Andreas Nugaard Holm, Barbara Plank, Dustin Wright, and Isabelle Augenstein. Longitudinal citation prediction using temporal graph neural networks. In *SDU@AAAI*, 2022. DBLP License: DBLP's bibliographic metadata records provided through http://dblp.org/ are distributed under a Creative Commons CC0 1.0 Universal Public Domain Dedication. Although the bibliographic metadata records are provided consistent with CC0 1.0 Dedication, the content described by the metadata records is not. Content may be subject to copyright, rights of privacy, rights of publicity and other restrictions.

Alfonso Ibáñez, Pedro Larrañaga, and Concha Bielza. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309, 2009.

Song Jiang, Bernard Koch, and Yizhou Sun. Hints: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *Proceedings of the web conference 2021*, pp. 3158–3167, 2021.

Yiqiao Jin, Yijia Xiao, Yiyang Wang, and Jindong Wang. Scito2m: A 2 million, 30-year cross-disciplinary dataset for temporal scientometric analysis. In *Workshop on Preparing Good Data for Generative AI: Challenges and Approaches*, 2024.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time, 2019. URL https://arxiv.org/abs/1907.05321.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

Vincent Larivière and Yves Gingras. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1):126–131, 2010.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Charles Spearman. The proof and measurement of association between two things. 1961.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4 (11):992–1003, 2011.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990–998, 2008.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Adilson Vital Jr, Filipi N Silva, Osvaldo N Oliveira Jr, and Diego R Amancio. Predicting citation impact of research papers using gpt and other text embeddings. *Physica A: Statistical Mechanics and its Applications*, 674:130789, 2025.

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

Wanjun Xia, Tianrui Li, and Chongshou Li. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, 128(1):543–585, 2023.

Xovee Xu, Ting Zhong, Ce Li, Goce Trajcevski, and Fan Zhou. Heterogeneous dynamical academic network for learning scientific impact propagation. *Knowledge-Based Systems*, 238:107839, 2022.

Zhikai Xue, Guoxiu He, Zhuoren Jiang, Sichen Gu, Yangyang Kang, Star Zhao, and Wei Lu. Predicting scientific impact through diffusion, conformity, and contribution disentanglement. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 2764–2774, 2024.

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

Carl Yang and Jiawei Han. Revisiting citation prediction with cluster-aware text-enhanced heterogeneous graph neural networks. In *2023 IEEE 39th international conference on data engineering (ICDE)*, pp. 682–695. IEEE, 2023.

Fang Zhang and Shengli Wu. Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics*, 129(7):4137–4166, 2024.

Kai Zhang, Kaisong Song, Yangyang Kang, and Xiaozhong Liu. Content-and topology-aware representation learning for scientific multi-literature. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7490–7502, 2023.

Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. From words to worth: Newborn article impact prediction with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1183–1191, 2025.

Xin Ping Zhu and ZhiJie Ban. Citation count prediction based on academic network features. In *2018 IEEE 32nd international conference on advanced information networking and applications (AINA)*, pp. 534–541. IEEE, 2018.

Appendix

Table of Contents

A Dataset Construction and Statistics **13**
A.1 Metadata Statistics 14
A.2 Split Dataset Distribution 15

B Loss Function Details **16**

C Evaluation Metrics Definition **17**
C.1 Year-Level Metrics 17
C.2 Month-Level Metrics 18

D More Experimental Results **18**
D.1 Significance Test 18
D.2 More Case Study 18
D.3 Domain Performance Differences 19

E TNCSI Calculation **19**

F Training parameters **19**

G Prompt **21**

H Limitations and Future Work **27**

A Dataset Construction and Statistics

Comparison with Existing Datasets. Table 6 situates our benchmark in the landscape of scientific paper datasets. Existing resources each contribute valuable aspects: APS (Bai et al., 2019) and DBLP (Zhu & Ban, 2018) provide structured networks but lack multimodality; PubMed (Fu & Aliferis, 2008) offers scale but limited label granularity; S2AG [1] achieves massive coverage but with sparse annotations; Aminer [2] and Bio-Sci (Zhang et al., 2023) capture recent slices but often full content. Our dataset is, to our knowledge, the first at this scale (2.0M query papers, 39.3M-node ego-graph) to unify multi-grained impact labels, real author/venue/institution metadata, and built-in processing tools within a single resource – while being explicit, rather than silent, about the domain shift an unstratified sample of this kind inevitably carries (see Table 1).

Table 6: Comparison of popular impact prediction paper datasets and databases. Abbreviations: Gran. = multi-grained (year- and month-level) labels, Dom. = multi-domain coverage. †Our domain coverage is the *natural*, unstratified distribution of eligible S2AG papers, not a balanced sample – see Table 1 for the resulting shift relative to the full corpus.

Dataset	Year	Gran.	Dom.	Tools	#Papers
APS	–	✗	✗	✗	0.45M
Pubmed	2008	✗	✗	✗	1.10M
DBLP	2008	✗	✗	✗	1.80M
S2AG	2022	✓	✓	✗	205M
Aminer	2023	✗	✗	✗	5.25M
Bio-Sci	2023	✗	✓	✗	0.03M
Ours	2026	✓	✓†	✓	2.0M

[1] https://www.semanticscholar.org/
[2] https://www.kaggle.com/datasets/kmader/aminer-academic-citation-dataset

Table 7: Full corpus vs. sampled query set, computed over `paper_static_features` (full corpus) and the union of `splits.parquet`/`node_features.parquet` (sample). Field shares are multi-label and computed independently within each population; ranks refer to each population's own top-15 by share. Full 15-field ranking in Appendix A, Table 8.

Attribute	Full S2AG corpus	Sampled query set
# Papers	237,026,483	2,000,000 (0.84%)
Publication years	1970–2025	2005–2020
Zero-citation share	53.4%	13.8%
5,000+-citation share	0.003–0.008%	0.003–0.008%
Physics (share, rank)	9.7%, 8th	21.4%, 3rd
History / Political Science in top-15	✓ (14th, 11th)	✗ (dropped)

A.1 Metadata Statistics

Source database. All papers, authors, venues, affiliations, and citation edges are drawn from a single Semantic Scholar Academic Graph (S2AG) snapshot – a 307GB database of 237,026,483 papers spanning 1970–2025. Citation structure comes from the `citations_by_cited`/`citations_by_citing` edge tables (5.7B rows each; every real citation edge is stored twice in this snapshot, confirmed against `papers.citationcount` and the live Semantic Scholar API, so every query in our pipeline de-duplicates on `(citing_corpusid, cited_corpusid)` before counting), authorship from `paper_authors` (697.7M raw rows), and affiliations from `paper_affiliations`. We additionally cross-checked S2AG's own `fields_of_study` classification coverage over time: it holds around 65% for papers published before 2021 and drops to roughly 32% from 2021 onward without recovering in this snapshot, which is one reason (together with needing a full 5-year label window) that our query window stops at 2020.

Domain distribution. Table 8 shows the top-15 fields-of-study by share within our 2,000,000-query sample (a paper may carry more than one field tag, so shares do not sum to 100%). The distribution is *not* uniform: Medicine alone tags 75.6% of query papers, roughly triple the next-largest field (Biology, 26.6%), and coverage falls off from there through Computer Science (21.2%), Engineering (20.9%), down to Education (3.5%) at 15th. This is an honest consequence of unstratified sampling over S2AG's own field mix, not a sampling target – Table 7.

Table 8: Top-15 field-of-study ranking, full S2AG corpus vs. our sampled query set. History and Political Science (full-corpus ranks 14 and 11) fall out of the sample's top-15 entirely, replaced by Agricultural and Food Sciences and Education – consistent with the referencecount/abstract eligibility filter retaining STEM papers at a higher rate than humanities/social-science papers.

Rank	Full S2AG corpus (share)	Sampled query set (share)
1	Medicine (42.9%)	Medicine (75.6%)
2	Engineering (15.8%)	Biology (26.6%)
3	Computer Science (12.7%)	Physics (21.4%)
4	Biology (12.5%)	Computer Science (21.2%)
5	Chemistry (11.2%)	Engineering (20.9%)
6	Environmental Science (10.7%)	Environmental Science (18.6%)
7	Materials Science (9.9%)	Materials Science (16.3%)
8	Physics (9.7%)	Chemistry (14.2%)
9	Psychology (5.7%)	Mathematics (12.1%)
10	Business (4.4%)	Psychology (8.6%)
11	Political Science (4.4%)	Economics (5.0%)
12	Mathematics (4.3%)	Sociology (4.9%)
13	Sociology (4.0%)	Business (4.8%)
14	History (3.9%)	Agricultural and Food Sciences (3.6%)
15	Economics (3.7%)	Education (3.5%)

Year distribution, overall and by split. Figure 5 shows the publication-year share of the full 2M-query set (gray bars) against each individual split (colored lines). Because `train`, `val`, and `previous_test` are all drawn without overlap from the same 2005–2017 historical pool via the same random partition, their year distributions are visually indistinguishable from one another – all three trace the same growing-then-plateauing curve, rising from 2.1–2.2% of the split in 2005 to a peak of 15.6–15.9% in 2017 (more eligible papers survive the referencecount/abstract filter in later years). `fresh_test`, by contrast, is drawn entirely from 2018–2020 and is a disjoint block in publication-year space by construction, not merely by chance – confirming the split does what Section 3 claims: `fresh_test` contains publication years the model never saw during training.

A.2 Split Dataset Distribution

Figure 6 compares the golden citation labels themselves across splits: the left panel plots the median cumulative year-level citation count at each horizon k=1..5, and the right panel the mean cumulative month-level count (mean rather than median, since Figure 2a already shows the month-level median is 0 at every horizon). `train`, `val`, and `previous_test` again track each other almost exactly at both granularities (e.g. year-5 median of 5.0 citations, month-5 mean of 0.43), consistent with being three random draws from the same underlying pool. `fresh_test` sits systematically higher at every horizon and every granularity – year-5 median 6.0 vs. 5.0, month-5 mean 0.68 vs. 0.43 – which is expected rather than a sampling artifact: 2018–2020 papers are, on average, in more actively-growing subfields of a corpus whose overall publication volume keeps rising, so a same-length citation window started more recently tends to land at a higher point on that growth curve. This also means a model trained on `train` (drawn from the older, lower-volume regime) faces a real, systematic level shift when evaluated on `fresh_test` – not just a different time window, but a different citation-rate regime – which is precisely the generalization gap `fresh_test` is designed to expose.

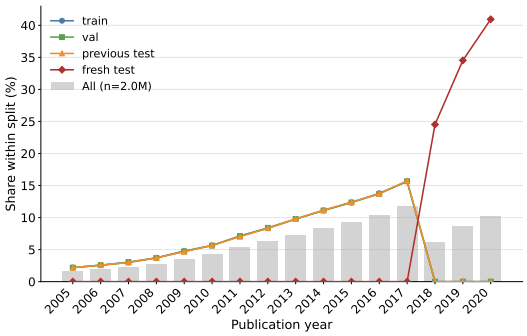


Figure 5: Publication-year share of the full query set (gray bars) and of each split (colored lines). `train`/`val`/`previous_test` overlap almost exactly (same 2005–2017 pool, same random partition); `fresh_test` occupies a disjoint 2018–2020 block.

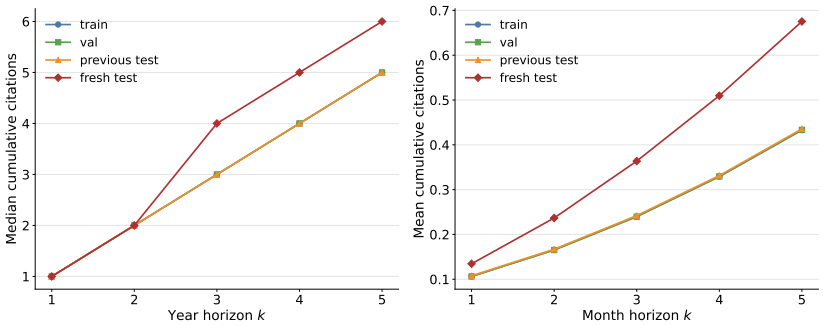


Figure 6: Golden-label citation trajectories by split. **Left:** median cumulative year-level citations at each horizon. **Right:** mean cumulative month-level citations at each horizon (mean used since the month-level median is 0 everywhere, per Figure 2a). `train`/`val`/`previous_test` overlap almost exactly; `fresh_test` is systematically higher at every horizon and both granularities.

B Loss Function Details

The composite objective in Eq. 6 combines a primary regression loss with three auxiliary terms, each targeting a different part of the architecture.

Regression loss $\mathcal{L}_{\text{reg}}$. Given the fused, calibrated prediction $\hat{y} \in \mathbb{R}^T$ for the T future citation horizons and the ground-truth cumulative citation counts y, we minimize mean squared error in $\log(1+x)$ space:

$$\mathcal{L}_{\text{reg}} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t - \log(1+y_t)\right)^2, \tag{7}$$

averaged over the batch. When `--group_reweight` is enabled (off by default), this per-sample MSE is instead combined with learned domain- and time-specific log-variance terms $s_d = \texttt{domain_log_var}[d]$, $s_\tau = \texttt{time_log_var}[\tau]$ in a Kendall-style homoscedastic uncertainty weighting, $\mathcal{L}_{\text{reg}} = e^{-(s_d+s_\tau)} \cdot \text{MSE} + s_d + s_\tau$, so the model can automatically down-weight noisy domain/time slices; we did not enable this by default and report results with the plain MSE above.

Citation-bucket auxiliary loss $\mathcal{L}_{\text{bucket}}$. Each paper is assigned one of three coarse citation-impact buckets based on its 5-year cumulative citation count c_5: *low* ($c_5 < 10$), *medium* ($10 \le c_5 < 100$), and *high* ($c_5 \ge 100$). A linear classification head reads the fused branch embedding and predicts bucket logits $\ell \in \mathbb{R}^3$; $\mathcal{L}_{\text{bucket}}$ is the standard cross-entropy against the ground-truth bucket label. This is an auxiliary supervision signal only – it does not participate in the final citation-count prediction – intended to encourage the fused representation to also be linearly separable by coarse impact tier.

Contrastive loss $\mathcal{L}_{\text{contrast}}$. To encourage the graph branch's pooled embedding to be robust to perturbations of the citation ego-graph's neighborhood, we apply a hard-negative InfoNCE loss following H2CGL (He et al., 2023). For each anchor graph we construct two augmented views by randomly dropping a fraction ρ (`--aug_rate`, default 0.1) of droppable (non-target, non-venue) neighbor nodes, biased toward dropping low-popularity nodes for the "positive" view z_1 and high-popularity nodes for the mined-hard-negative view; batch-mined hard negatives z^- (when available) are pooled with the in-batch negatives. With $\text{sim}(u,v) = u^\top v/\tau$ ($\tau = 0.2$) computed against the pool of all positives and negatives, excluding the anchor itself:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_1, z_2))}{\sum_{v \in \{z_2\}\cup\mathcal{N}\cup\{z^-\}} \exp(\text{sim}(z_1, v))}, \tag{8}$$

where $\mathcal{N}$ is the set of in-batch negatives. On *full_dataset* we do not yet have a mined hard-negative pool, so $\lambda_{\text{cl}} = 0$ there and this term reduces to a standard in-batch InfoNCE (or is fully disabled); on *calib_50k* the full hard-negative-augmented version is used.

Branch deep-supervision loss $\mathcal{L}_{\text{branch}}$. Each active branch (LLM, graph, temporal, retrieval, track-record, explicit- feature, as configured) additionally predicts its own citation estimate $\hat{y}^{(b)}$ from its own embedding via a lightweight branch-specific head, independent of the gated fusion. We average the per-branch MSE (same $\log(1+x)$ target as $\mathcal{L}_{\text{reg}}$) over the B active branches:

$$\mathcal{L}_{\text{branch}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t^{(b)} - \log(1+y_t)\right)^2. \tag{9}$$

This keeps every branch directly regression-supervised even though only the fused prediction is the model's primary output, preventing a branch from degenerating into a component the gate can freely ignore during early training.

Weights and training settings. We set $\lambda_{\text{aux}} = 0.5$ (`--aux_weight`), $\lambda_{\text{cl}} = 0.5$ (`--cl_weight`, 0 on *full_dataset*), and $\lambda_{\text{branch}} = 0.2$ (`--branch_aux_weight`) for all reported results unless otherwise noted as an ablation.

C Evaluation Metrics Definition

We report RMSE for year-level prediction and AUROC, PR-AUC, QWK, RMSE, and NDCG@20 for month-level prediction in our main comparison tables (Table 2 and Table 3). Year-level prediction focuses on regression accuracy, while the month-level task additionally considers zero-inflation-aware classification and ranking metrics. Following prior work on citation forecasting (He et al., 2023; Xue et al., 2024), all regression targets are evaluated in $\log(1+x)$ space to reduce the influence of the heavy right tail of citation-count distributions. We consider the same set of $H = 5$ future citation horizons for both prediction granularities: years 1–5 for the *Year* target and months 1–5 for the *Month* target.

Notation. Let $\mathcal{D}_{\text{test}} = \{1, \ldots, N\}$ denote a test split, such as PREVIOUS_TEST or FRESH_TEST. Let $c_{i,h}$ denote the raw cumulative citation count of paper i at horizon $h \in \{1, \ldots, H\}$. We evaluate regression metrics in $\log(1+x)$ space:

$$y_{i,h} = \log(1 + c_{i,h}), \qquad \hat{y}_{i,h} = \log(1 + \hat{c}_{i,h}), \tag{10}$$

where $\hat{c}_{i,h}$ is the predicted cumulative citation count. For ranking and regression metrics, the transformed values are used consistently throughout evaluation.

C.1 Year-Level Metrics

RMSE. We report per-horizon RMSE,

$$\text{RMSE}_h = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{i,h} - y_{i,h})^2}, \tag{11}$$

and define the pooled RMSE across all $N \times H$ paper–horizon pairs as

$$\text{RMSE}_{\text{overall}} = \sqrt{\frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H}(\hat{y}_{i,h} - y_{i,h})^2}. \tag{12}$$

The *Year Avg* and *Month Avg* columns in Table 2 and Table 3 report the unweighted mean of the per-horizon RMSE values,

$$\text{RMSE}_{\text{Avg}} = \frac{1}{H}\sum_{h=1}^{H}\text{RMSE}_h. \tag{13}$$

This treats all prediction horizons equally and prevents horizons with larger residual variance from dominating the aggregate score.

NDCG@20. RMSE and classification metrics do not directly measure whether a model correctly ranks papers by future citation impact. We therefore report NDCG@20 to evaluate the quality of the highest-ranked papers. For each horizon h, we treat the entire test split as a single ranked list, sorting papers according to the model's predicted scores $\hat{y}_{\cdot,h}$. The discounted cumulative gain at rank k is

$$\text{DCG}_h@k = \sum_{r=1}^{k}\frac{2^{\text{rel}(r)} - 1}{\log_2(r+1)}, \tag{14}$$

where $\text{rel}(r)$ is the ground-truth relevance of the paper ranked at position r, here given by $y_{i,h}$. We set $k = 20$ and normalize by the ideal DCG obtained by ranking papers according to ground-truth relevance:

$$\text{NDCG}_h@20 = \frac{\text{DCG}_h@20}{\text{IDCG}_h@20}. \tag{15}$$

We report NDCG@20 as the mean of the five per-horizon values. Higher NDCG@20 indicates better agreement between the model's top-20 predicted ranking and the ranking induced by realized citation impact.

NDCG@20 is computed globally over the full evaluation split rather than independently within queries, batches, or distributed evaluation shards. When evaluation is parallelized across multiple devices, we first reassemble the predictions and ground-truth labels from all shards and then compute NDCG@20 once over the complete split. This avoids averaging rank-based scores computed on disjoint subsets, which would not preserve the ranking relationships in the full evaluation set.

C.2 MONTH-LEVEL METRICS

AUROC and PR-AUC. Month-level citation targets are highly zero-inflated, with 85.2% of the valid previous-test labels being zero. We therefore complement regression with binary discrimination metrics. For each horizon h, we define the binary citation indicator

$$z_{i,h} = \mathbb{I}[c_{i,h} > 0], \tag{16}$$

where $z_{i,h} = 1$ indicates that the paper receives at least one citation by horizon h. The model's predicted citation score is used as the continuous decision score for both metrics.

AUROC measures the probability that a randomly selected cited paper receives a higher predicted score than a randomly selected uncited paper. Equivalently, it is the area under the receiver operating characteristic curve obtained by varying the decision threshold over the predicted scores. PR-AUC is the area under the precision–recall curve and is particularly informative under severe class imbalance. We report the corresponding scores for each month-level horizon and aggregate them consistently across the five horizons.

Quadratic Weighted Kappa (QWK). To evaluate ordinal agreement between predicted and realized citation impact, we discretize cumulative citation counts into three ordered impact buckets, *low*, *medium*, and *high*, using fixed citation-count thresholds. The same thresholds are applied to all models and splits. Let $q_{i,h}$ and $\hat{q}_{i,h}$ denote the ground truth and predicted ordinal categories, respectively. QWK measures the agreement between the two ordinal labelings while assigning larger penalties to larger disagreements. Following the standard formulation,

$$\mathrm{QWK} = 1 - \frac{\sum_{a,b} w_{ab} O_{ab}}{\sum_{a,b} w_{ab} E_{ab}}, \tag{17}$$

where O_{ab} is the observed confusion matrix, E_{ab} is the expected confusion matrix under the empirical marginal distributions, and

$$w_{ab} = \frac{(a-b)^2}{(K-1)^2}, \qquad K = 3. \tag{18}$$

Higher QWK indicates stronger ordinal agreement between predicted and ground-truth citation-impact levels.

Splits. We report the evaluation metrics separately on PREVIOUS_TEST and FRESH_TEST. The PREVIOUS_TEST split contains papers whose publication years overlap with the training period and therefore evaluates in-distribution temporal generalization. In contrast, FRESH_TEST contains papers published strictly after the training cutoff and evaluates extrapolation under temporal distribution shift. This separation allows us to distinguish predictive performance within the historical data distribution from robustness to unseen publication periods.

D MORE EXPERIMENTAL RESULTS

D.1 SIGNIFICANCE TEST

We assess the statistical significance of the performance differences between *LLM4Impact* and its ablations using a paired bootstrap test. Predictions are matched by `corpusid`, ensuring that each comparison is performed on exactly the same set of papers. As shown in Table 9, we draw 2,000 bootstrap resamples and report the RMSE difference, defined as the ablation RMSE minus the RMSE of *LLM4Impact*, together with its 95% confidence interval and p-value. All comparisons yield $p < 0.0001$, with confidence intervals entirely above zero, confirming that the observed improvements of *LLM4Impact* over the ablated variants are statistically significant on both the previous and fresh test sets.

D.2 MORE CASE STUDY

To facilitate qualitative inspection of our method's behavior, we accompany this submission with an interactive demo website (included in the supplementary material) that allows a reader to explore model predictions on real papers without running any code. The demo provides four components, illustrated in Figures 8–12:

Table 9: Paired bootstrap significance tests comparing *LLM4Impact* with its ablations. RMSE differences are computed as $\mathrm{RMSE}_{\mathrm{abl.}} - \mathrm{RMSE}_{\textit{LLM4Impact}}$. Positive values indicate that the ablation performs worse than *LLM4Impact*.

Ablation	Split	RMSE Diff.	95% CI	p
no_graph	previous_test	+0.0482	[0.0467, 0.0496]	< 0.0001
no_llm	previous_test	+0.0891	[0.0872, 0.0910]	< 0.0001
no_llm_graph_prefix	previous_test	+0.0033	[0.0023, 0.0043]	< 0.0001
no_graph	fresh_test	+0.0328	[0.0312, 0.0343]	< 0.0001
no_llm	fresh_test	+0.1028	[0.1010, 0.1046]	< 0.0001

- **Case Gallery** (Figure 8): a curated set of real papers from top CS/ML venues, each showing our method's predicted citation trajectory alongside five baselines and the actual (golden) citation counts, together with the paper's full metadata (title, venue, publication year, authors, institutions, categories, abstract, and references). Cases are tagged as wins, typical, or losses for our method, so the gallery is not restricted to favorable examples.
- **Dataset Examples** (Figure 9 and 10): a browsable collection of papers from the evaluation pool with complete metadata and their ground-truth citation trajectories at both year and month granularity, illustrating the task itself independent of any model's predictions.
- **Surface-Cue Ablation Study** (Figure 11): an interactive view of the single-variable metadata perturbation study described in Section 5.4 (author, publication year, venue, and reference perturbations), with per-paper drill-down.
- **Content-Cue Ablation Study** (Figure 12): an interactive view of the abstract-content perturbation study, isolating what our LLM branch reads from the paper text itself.

We sincerely hope readers to consult the submitted website directly for additional examples beyond what is shown here.

D.3 Domain Performance Differences

Figure 7 breaks down the previous_test/fresh_test gap by primary field of study. The gap is not uniform across domains: **Computer Science shows both the largest absolute RMSE and the largest previous_test→fresh_test gap** (+0.13), suggesting its citation dynamics are both harder to predict overall and less stable across time periods than other fields. In contrast, **History has an essentially zero gap** (-0.00), and several small-sample humanities domains (Law, Philosophy, Linguistics) also show gaps under $+0.05$, consistent with citation patterns in these fields evolving more slowly and thus generalizing better to a future test period. This pattern held across all 22 domains with $\geq 1{,}000$ examples in our fresh_test split.

E TNCSI Calculation

TNCSI (Topic Normalized Citation Success Index) is a metric designed to evaluate the citation impact of articles by comparing them within the same research field, providing a normalized score between 0 and 1 that indicates the likelihood an article's impact surpasses that of its peers. Unlike traditional citation counts, TNCSI normalizes across fields but originally focused on review papers and cumulative citations, limiting its suitability for comparing newly published or regular research articles. The formulation is as follows:

$$TNCSI = \int_0^{\text{cites}} \lambda e^{-\lambda x}, dx, \quad x \geq 0. \tag{19}$$

F Training parameters

We implement all branches in PyTorch. Our main results use `Qwen3-0.6B` as the frozen LLM backbone; for a same-backbone comparison against NAIP, we additionally report a `Qwen2.5-7B-Instruct` backbone variant. Both are adapted with LoRA ($r = 64$, $\alpha = 32$,

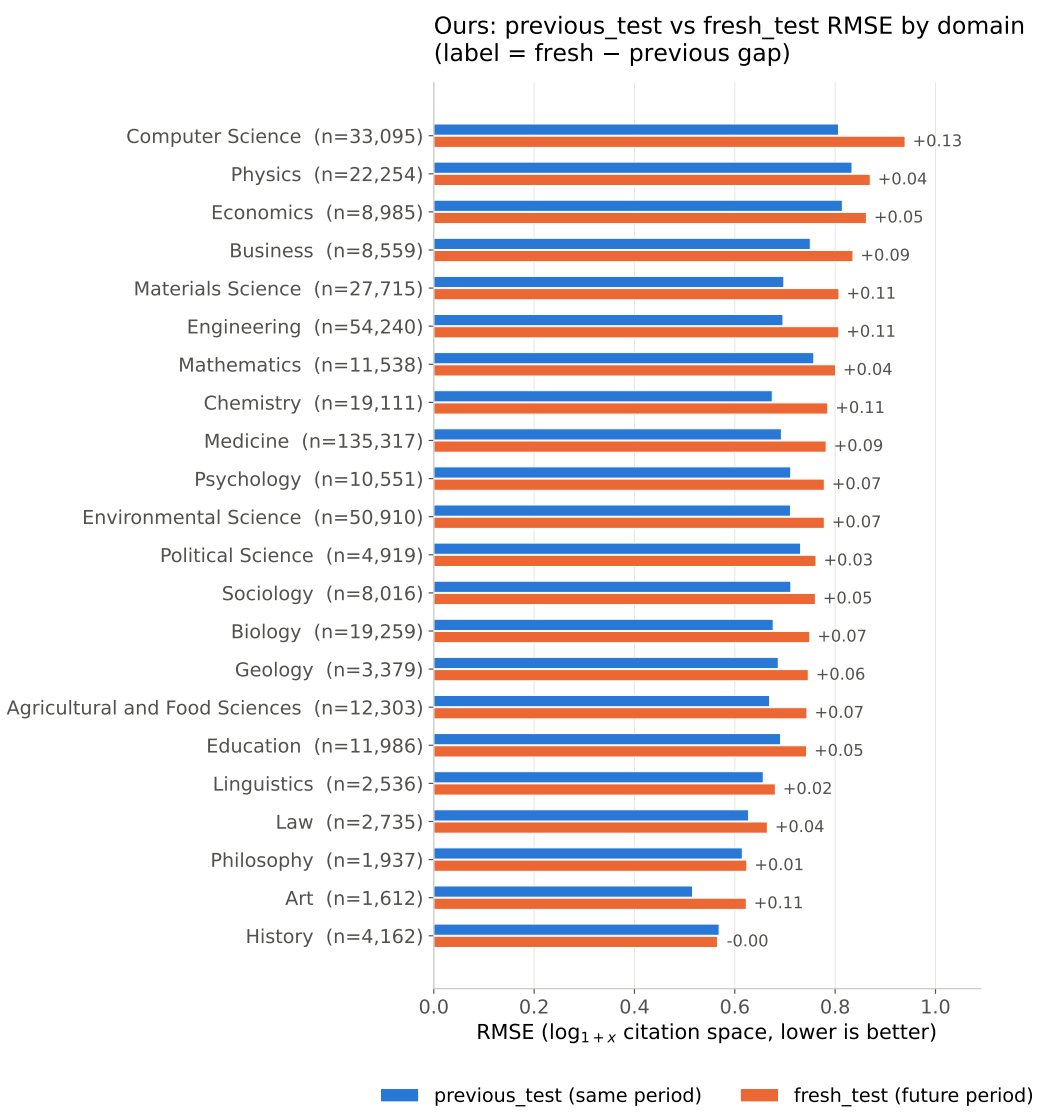
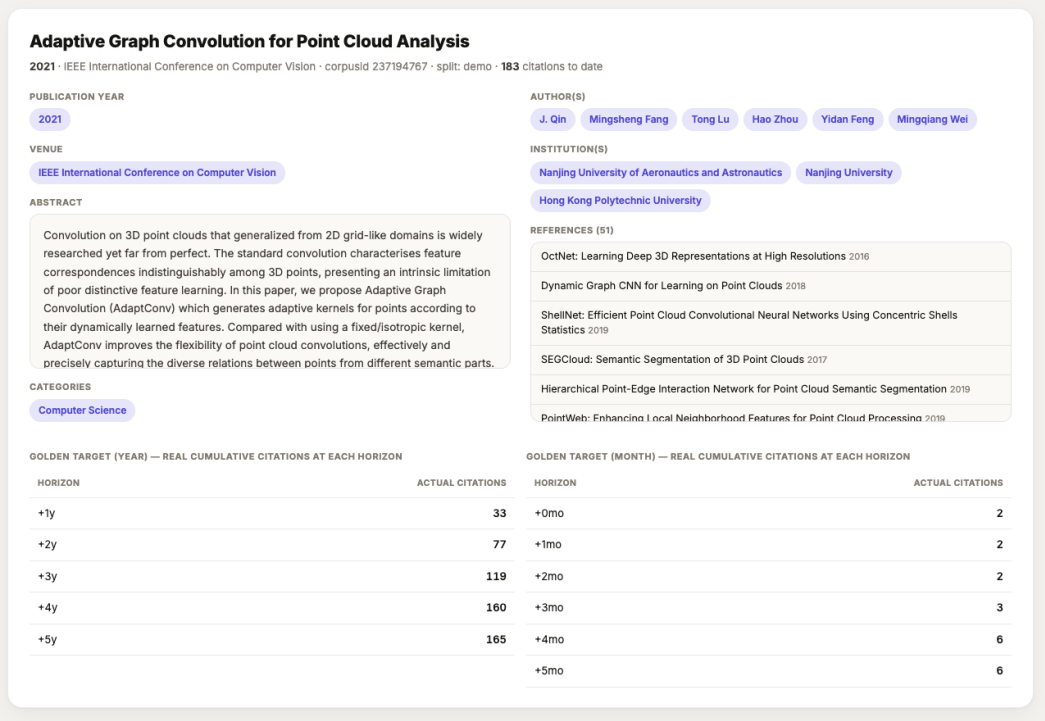


Figure 7: RMSE of the full model on *previous_test* (same time period as training) versus *fresh_test* (held-out future time period), broken down by primary field of study. Domains are sorted by *fresh_test* RMSE (easiest at bottom). The label at the end of each domain's bars is the fresh−previous gap, i.e. how much harder extrapolating to a future time period is for that field.

applied only to the attention $q_{\text{proj}}/v_{\text{proj}}$ projections) rather than full fine-tuning. The graph branch encodes a 1-hop ego-tree around each target paper with a 3-layer, 8-head relational graph attention network, and its pooled embedding is projected into 10 soft prefix tokens prepended to the LLM's input sequence. We optimize the composite objective in Eq. 6 with AdamW at a constant learning rate of 2×10^{-5} (no warmup or decay schedule) and weight decay 0.01, training for a single full epoch over the 800K-example training set on $4\times$ NVIDIA A100-40GB GPUs under distributed data parallelism, with validation checked every 500 steps (used only for checkpoint selection, not for early termination – we always complete the full epoch). Per-GPU batch size is 8 (effective 32) for the 0.6B backbone and 4 (effective 16) for the 7B backbone, reduced to fit the larger backbone's memory footprint alongside the graph branch. We evaluate on two disjoint test splits: *previous_test* (400K examples, same time period as training) and *fresh_test* (500K examples, a held-out future time period), predicting cumulative citation counts at $T = 5$ future horizons. Table 10 summarizes the main hyperparameters; the auxiliary contrastive loss weight λ_{cl} is set to 0.5 on the full dataset.



Adaptive Graph Convolution for Point Cloud Analysis

2021 · IEEE International Conference on Computer Vision · corpusId 237194767 · split: demo · **183** citations to date

PUBLICATION YEAR: 2021

AUTHOR(S): J. Qin, Mingsheng Fang, Tong Lu, Hao Zhou, Yidan Feng, Mingqiang Wei

VENUE: IEEE International Conference on Computer Vision

INSTITUTION(S): Nanjing University of Aeronautics and Astronautics, Nanjing University, Hong Kong Polytechnic University

ABSTRACT

Convolution on 3D point clouds that generalized from 2D grid-like domains is widely researched yet far from perfect. The standard convolution characterises feature correspondences indistinguishably among 3D points, presenting an intrinsic limitation of poor distinctive feature learning. In this paper, we propose Adaptive Graph Convolution (AdaptConv) which generates adaptive kernels for points according to their dynamically learned features. Compared with using a fixed/isotropic kernel, AdaptConv improves the flexibility of point cloud convolutions, effectively and precisely capturing the diverse relations between points from different semantic parts.

REFERENCES (51)

- OctNet: Learning Deep 3D Representations at High Resolutions 2016
- Dynamic Graph CNN for Learning on Point Clouds 2018
- ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics 2019
- SEGCloud: Semantic Segmentation of 3D Point Clouds 2017
- Hierarchical Point-Edge Interaction Network for Point Cloud Semantic Segmentation 2019
- PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing 2019

CATEGORIES: Computer Science

GOLDEN TARGET (YEAR) — REAL CUMULATIVE CITATIONS AT EACH HORIZON

HORIZON	ACTUAL CITATIONS
+1y	33
+2y	77
+3y	119
+4y	160
+5y	165

GOLDEN TARGET (MONTH) — REAL CUMULATIVE CITATIONS AT EACH HORIZON

HORIZON	ACTUAL CITATIONS
+0mo	2
+1mo	2
+2mo	2
+3mo	3
+4mo	6
+5mo	6

Figure 8: Example benchmark instance and data format, including paper metadata, references, and ground-truth citation targets at year and month granularities.

Table 10: Main hyperparameters for the reported full-dataset results.

Category	Hyperparameter	Value
Architecture	Hidden dimension	256
	Graph layers / attention heads	3 / 8
	LLM backbone	Qwen3-0.6B / Qwen2.5-7B-Instruct[‡]
	LoRA rank r / alpha α / dropout	64 / 32 / 0.05
	Graph-prefix length	10 tokens
Loss weights	λ_{aux} (bucket classification)	0.5
	λ_{cl} (contrastive)	0.5
	λ_{branch} (deep supervision)	0.2
Optimization	Optimizer	AdamW
	Learning rate (constant)	2×10^{-5}
	Weight decay	0.01
	Per-GPU / effective batch size	8 / 32 (0.6B), 4 / 16 (7B)
	GPUs	4×A100-40GB
	Training length	1 full epoch
Data	Train / val examples	800K / 300K
	previous_test / fresh_test examples	400K / 500K
	Prediction horizons T	5
	Ego-tree hop depth	1

G PROMPT

The prompt of LLMs to predict future citation counts.

You are a scientometrics analyst. The tokens preceding this text encode structural context about the paper – its citation graph, retrieved related papers, and its own venue, institution, authors, and references. Using both that context and the paper's own content below, predict its cumulative citation counts at several future time points after publication. Title: title Abstract: abstract

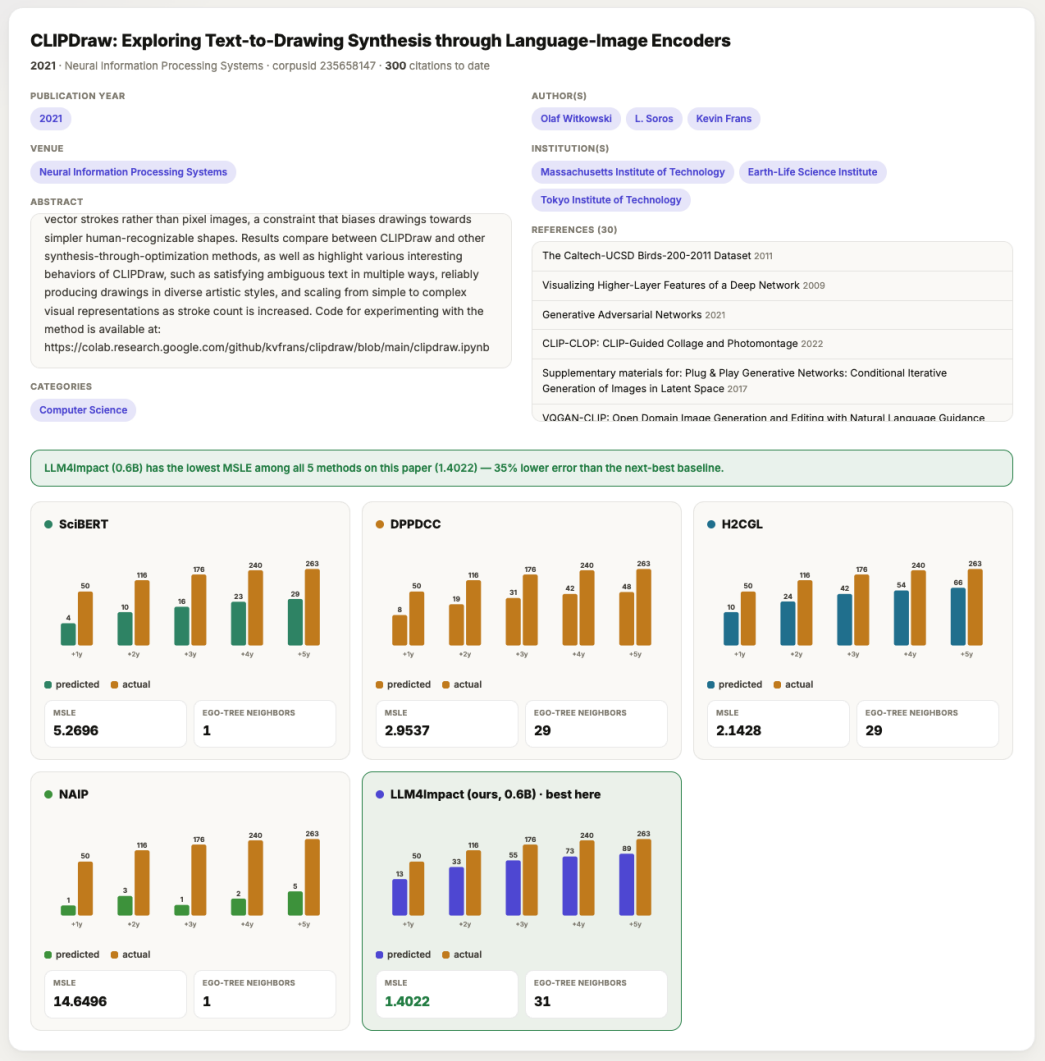


Figure 9: Interactive case study comparing LLM4Impact with baseline models on a real paper, showing predicted and ground-truth citation trajectories.

Table 11: Comparison of RMSE for raw citation counts and TNCSI, together with their Spearman correlation, on the previous test set. $C_{\log}$ denotes log-transformed citation counts and C_{TNCSI} denotes TNCSI-based impact scores.

Metric	Year_1	Year_2	Year_3	Year_4	Year_5	Avg
			Previous			
C_log	0.6772	0.6932	0.7809	0.8022	0.9177	0.7719
C_TNCSI	0.2016	0.2211	0.2291	0.2327	0.2352	0.2239
Cor.	0.7900	0.8061	0.8062	0.8119	0.8135	0.8055

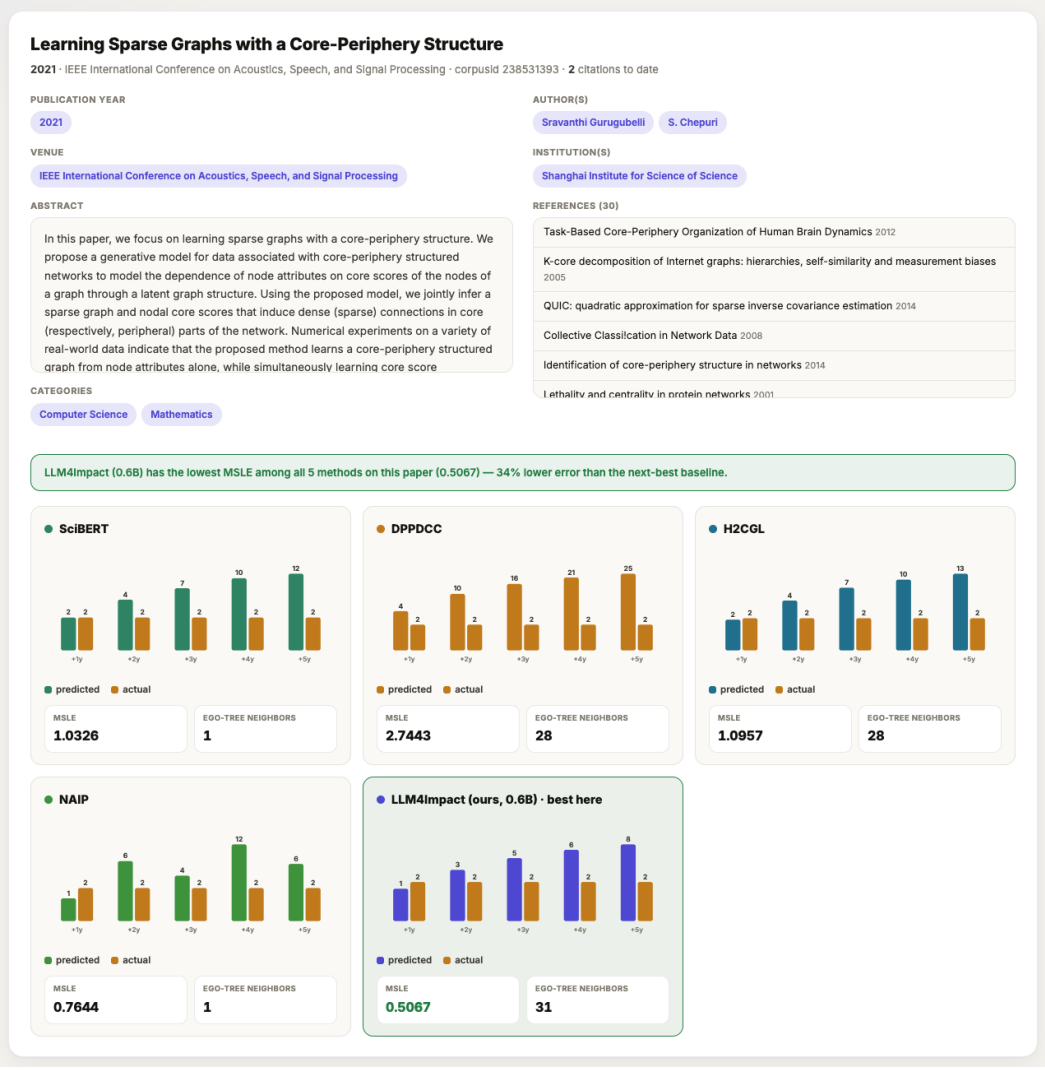


Figure 10: Second interactive case study comparing LLM4Impact with baseline models on a real paper, showing predicted and ground-truth citation trajectories.



Figure 11: Surface-cue perturbation study examining the effects of author, publication-year, venue, and reference changes.

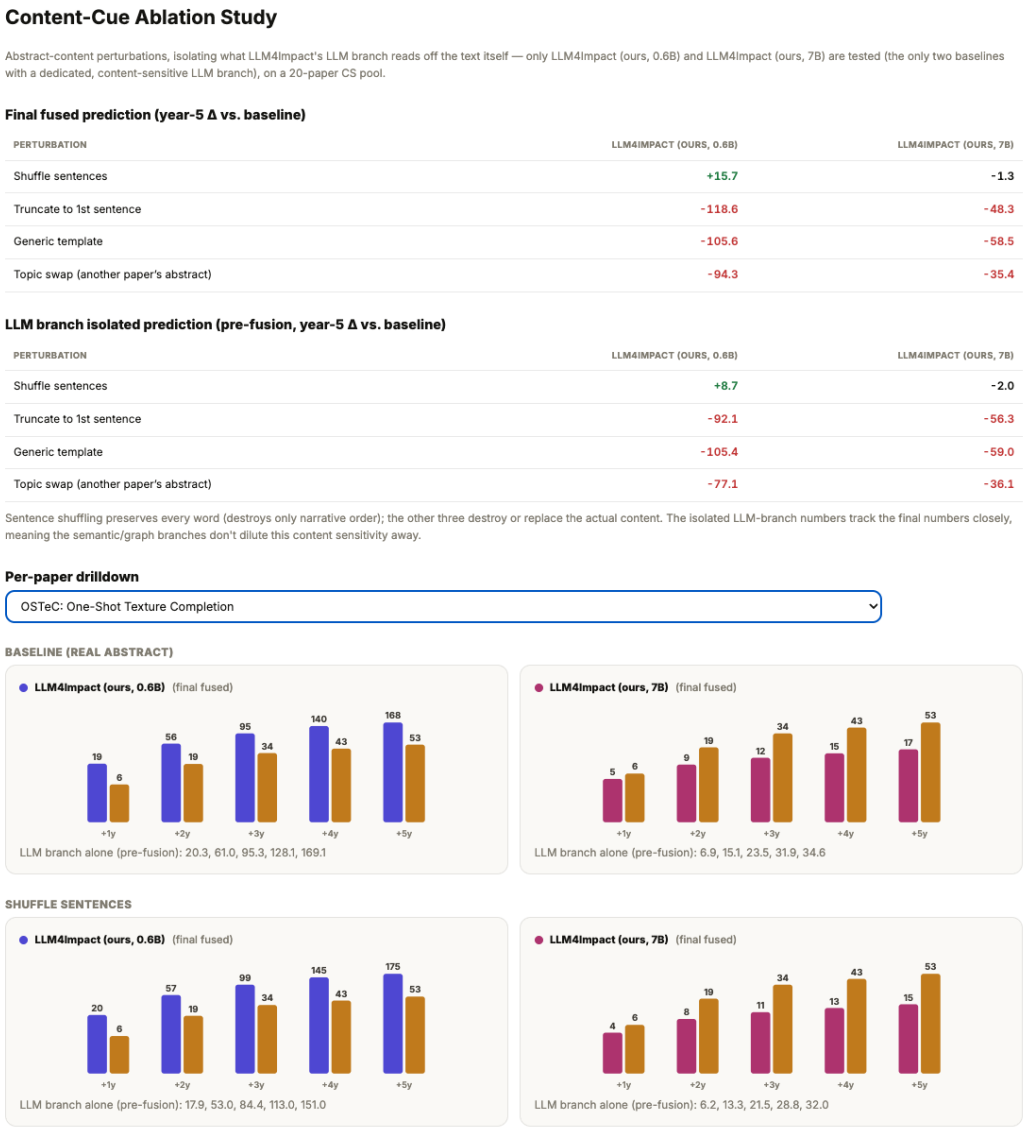


Figure 12: Content-cue perturbation study illustrating the role of the LLM branch under controlled abstract perturbations.

The prompt of LLMs to normalize noisy affiliations.

You are given a list of affiliation strings. Some are valid institutions, some are duplicated, and some are meaningless. Please normalize them as follows:

- If the string is a valid institution name but with formatting issues, fix it (e.g., remove extra punctuation, unify into "Institution Name").
- If the string is a duplicate of another name, map it to the same corrected name.
- If the string is not an institution name (number, meaningless), map it to "unknown".
- Return the result strictly as a JSON dictionary, where each original string is mapped to its corrected normalized name.

Example Input:
["University of California, Los Angeles", "University of California, Los Angeles", "01 Collaboration"]

Example Output:

```
{
  "University of California, Los Angeles": "University of
      California",
  "01 Collaboration": "unknown"
}
```

Generation Output:
Please normalize the following affiliation list accordingly.

The prompt of LLMs to normalize noisy venues.

You are given a list of venue strings. Some are valid venues, some are duplicated, and some are meaningless. Please normalize them as follows:

- If the string is a valid venue name but with formatting issues, fix it (e.g., remove extra punctuation, use proper title case like "Venue Name").
- Remove any explicit year tokens (e.g., "2015", "2014").
- Remove edition indicators such as ordinal numbers or words (e.g., "1st", "2nd", "Third", "Fourth").
- If the venue contains a trailing parenthetical acronym/abbreviation like "(ICAC3N)", "(HPEC)", "(IVAPP)" that is all caps/digits/hyphen and $\leq$10 chars, drop that parenthetical part. Keep descriptive tracks in parentheses (e.g., "(Emerging Technologies)") as is.
- If both a full spelled-out name and an acronym appear together, keep only the full spelled-out name and drop the acronym.
- If the string is a duplicate of another (after normalization), map it to the same corrected name.
- If the string is not a venue name (pure number or meaningless), map it to "unknown".
- Return the result strictly as a JSON dictionary, where each original string is mapped to its corrected normalized name. Do not include comments or extra text.

Example Input:
["Journal of Machine Learning Research", "2015 IEEE International Conference on Web Services", "4th International Conference on X (ICACX)", "01 Collaboration"]

Example Output:

```
{
  "Journal of Machine Learning Research": "Journal of Machine
      Learning Research",
  "2015 IEEE International Conference on Web Services": "IEEE
      International Conference on Web Services",
  "4th International Conference on X (ICACX)": "International
      Conference on X",
  "01 Collaboration": "unknown"
}
```

Generation Output:
Please normalize the following venue list accordingly.

H Limitations and Future Work

Our study focuses on a finite set of scientific domains, publication years, and prediction horizons, leaving broader settings for future investigation. In particular, longer-term citation dynamics and the continuous evolution of scholarly networks may provide additional signals beyond our current point-in-time representation. Future work will therefore explore broader domains, longer prediction horizons, and dynamic representations of evolving scientific knowledge.