Predicting Scholarly Impact with Temporal Preference Alignment

PARHAM HAMOUNI, Toronto Metropolitan University, Canada
EBRAHIM BAGHERI, University of Toronto, Canada

Background. Predicting the future influence of scientific papers remains a longstanding challenge in bibliometrics and information retrieval. Traditional regression and embedding-based methods estimate citation counts from textual features but fail to capture the inherently *relational and temporal* nature of scholarly impact.
Objective. This paper proposes a *preference-aligned framework* for forecasting and generating scholarly influence from paper abstracts and temporal cues. Rather than predicting absolute citation values, we model the relative likelihood that one paper will accrue more citations than another within a shared temporal time frame.
Methodology. Impact-DPO integrates temporally informed prompting with *direct preference optimization*, enabling LLMs to learn comparative influence patterns without explicit graph message passing. We formalize citation forecasting as pairwise preference learning on temporal text-attributed graphs, using publication year as a minimal temporal signal. Experiments were conducted on two large-scale domains, i.e.,*Computer Science* and *Physics*, spanning three temporal splits (2018–2020).
Results. We show that Impact-DPO achieves the highest pairwise accuracy across all splits, reaching up to *83%* outperforming SPECTER2+SVR by *12 pp* and zero-shot prompting by more than *15 pp*. Preference alignment yields an average improvement of roughly *+30 pp* over binary-classification baselines overall; importantly, supplementary same-backbone Qwen 2.5–7B BC runs remain near chance (about 50–58% across splits), showing that the gain is not explained by backbone capacity alone. A lower regularization parameter ($\beta = 0.1$) consistently produces optimal results, consistent with a low-gap preference regime in citation data. Generative evaluations further reveal that model-generated text aligns more closely with highly cited papers (KS=0.0744, p=0.0065), demonstrating emergent *generative alignment* with influential scholarly language.
Code and data availability. All code, data-preprocessing scripts, and evaluation notebooks are available at `impact-dpo` repository.

ACM Reference Format:
Parham Hamouni and Ebrahim Bagheri. 2025. Predicting Scholarly Impact with Temporal Preference Alignment. 1, 1 (July 2025), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Forecasting the future influence of information artifacts such as research papers, patents, or social media posts is a long-standing problem in information retrieval, scientometrics, and network science [6, 32, 57, 67]. The ability to anticipate which works will attract attention has both practical and epistemic importance. It informs funding priorities, peer review and recommendation systems, and the early identification of emerging research directions [51, 58, 70, 72]. Within scientific communication, accurate prediction of future impact enables more equitable recognition of novel

Authors' Contact Information: Parham Hamouni, parham.hamouni@torontomu.ca, Toronto Metropolitan University, Toronto, Ontario, Canada; Ebrahim Bagheri, ebrahim.bagheri@utoronto.ca, University of Toronto, Toronto, Ontario, Canada.

ACM XXXX-XXXX/2025/7-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

contributions and supports the stewardship of research ecosystems [6, 51, 70, 71]. Yet, despite decades of bibliometric research, predicting influence remains a challenging and largely unsolved problem. Influence evolves through dynamic and interdependent processes where textual contents signal intellectual novelty, network structure mediate visibility and diffusion, and temporal contexts determine when ideas resonate with the broader audience [18, 30, 58, 60, 63, 67]. As publication volumes surge and disciplinary boundaries shift, traditional predictors based on citation counts or static graph features have become increasingly brittle. A model of scientific influence must therefore reason across time, text, and structure simultaneously [6, 14, 57, 67, 71].

A fundamental challenge in this domain is the extent to which citation accumulation reflects intrinsic scientific quality versus extrinsic factors such as visibility, social influence, or cumulative advantage. Wang et al. [63] showed that long-term citation dynamics combine a preferential-attachment (rich-get-richer) component with an intrinsic fitness (quality) component, implying that raw citation counts conflate luck and visibility with genuine impact. Salganik et al. [50] demonstrated experimentally that social influence amplifies initial inequalities and makes success largely unpredictable, underscoring that popularity metrics reflect path-dependent dynamics as much as intrinsic quality. More broadly, Tahamtan and Bornmann [59] reviewed over a decade of citing-behavior studies and concluded that citations are driven by a complex mix of quality-related and quality-unrelated factors including journal prestige, author visibility, field size, and self-citation. Bao and Teplitskiy [1] further showed through simulation that rhetorical citation practices—references included for persuasive rather than intellectual-influence reasons—redistribute attention in ways that decouple citation counts from direct intellectual contribution. Our work does not claim that citation counts are a perfect proxy for quality; rather, we treat *relative* citation ordering within co-cited pairs as a weaker but more defensible signal, since papers sharing a common citing context are subject to similar visibility and cumulative-advantage conditions, thereby partially controlling for extrinsic confounds.

It is also important to distinguish the present work from recent studies that use citation-derived metrics such as the disruption index [31] to characterize *types* of impact (consolidating versus disruptive). Our objective is orthogonal in that we predict *which* of two co-cited papers will accumulate more citations by a future horizon, not whether a paper is disruptive. The preference-learning formulation therefore complements disruption-based analyses by providing a predictive, text-grounded mechanism for ranking future influence, whereas disruption indices are retrospective structural diagnostics computed after citations have accrued. Because our observation horizon is fixed at T=2022, the 2018–2019 cohorts are closer to medium-to-long-term forecasting, whereas the 2020 cohort should be interpreted as a medium-term test of influence ordering.

Within information science, the problem of forecasting scholarly impact lies at the intersection of temporal graph representation learning, large-language-model-based text understanding, and scientific impact modeling [6, 7, 57, 67, 71]. Temporal Text-Attributed Graphs (TAGs) provide a formal abstraction in which nodes represent documents with textual attributes and edges represent time-stamped relationships such as citations [5, 28, 61, 73]. Modeling these graphs requires capturing both semantic evolution and structural change over time. This task is difficult for several reasons. Temporal graph neural networks can track evolving topology but often underutilize the rich semantics contained in text and are computationally heavy at inference [6, 23, 27, 61]. Conversely, transformer-based language models capture linguistic and conceptual nuance but ignore the relational context in which ideas propagate [22, 34, 76]. Furthermore, citation networks exhibit strong non-stationarity where the same thematic or stylistic signals may correlate with success in one period and fade in another [18, 25, 38, 39, 41]. Effective prediction therefore demands temporal awareness, semantic sensitivity, and structural reasoning in a unified framework, which is currently an open challenge across graph learning and bibliometric modeling [6, 23, 61, 71].

1.1 Objectives and Direction

Against this backdrop, we focus on the task of *pairwise future-citation prediction* in temporal text-attributed graphs. Given a citing paper at time t and two of its referenced papers, the objective is to decide which of the two will accumulate more citations by a later horizon $T > t$ [14, 36, 45]. This formulation reframes influence prediction as a relative decision rather than an absolute regression problem. It aligns with how impact is typically assessed—through comparison—and reduces sensitivity to global citation inflation or disciplinary differences across years [41, 67]. Pairwise modeling also captures the intuition that influence is *context-dependent* where two papers competing for attention within the same intellectual niche provide a more meaningful comparison than arbitrary global rankings.

We view influence prediction as a preference-learning problem over temporally contextualized text pairs. Impact-DPO trains a large language model using preference optimization [46] on preference pairs derived from a citation graph. The graph structure informs the training and evaluation pairs and is used to label which paper in each pair ultimately receives more citations. At both training and inference time, the model operates on textual content, i.e., abstracts of the citing and cited papers, supplemented with temporal metadata including the year of publication.

This formulation combines *temporal prompting* [12], which enables reasoning about temporal order and recency, with *preference alignment*, which directly optimizes the comparative judgment central to the task. By uniting these elements, the model captures how semantic and temporal signals jointly shape future influence while remaining scalable and interpretable. It requires no graph message passing at inference, making it applicable to large-scale bibliometric datasets and other dynamic text-attributed networks. The approach thus bridges the strengths of text-based LLMs and temporal graph reasoning within a single preference-aligned framework.

1.2 Contributions

This work makes the following contributions to the literature.

(1) It formalizes the pairwise future-citation prediction problem as a structured preference-learning task on temporal text-attributed graphs.
(2) It proposes a direct preference optimization framework that aligns large language models with graph-derived preferences while maintaining text-only inference enhanced by temporal prompts.
(3) It introduces domain-stratified benchmarks from the unarXive dataset [49] covering Computer Science and Physics between 2018 and 2020, designed with strict temporal splits and identical pairwise test sets.
(4) It provides extensive comparisons against strong baselines, including embedding-based regressors [9, 52], prompting-only LLMs [11], a released impact predictor [74], and a binary-classification variant of our model [21].
(5) It offers empirical and analytical evidence that preference alignment with temporal cues yields consistent gains across architectures and domains, and that these gains persist even for pairs with small citation differentials—behaviour consistent with a low-gap preference regime whose causal grounding is examined in Section 7.
(6) It provides a detailed interpretability analysis combining counterfactual text interventions, activation steering, citation-gap analysis, and case-based evidence, showing that the model's learned preferences are grounded in scientifically meaningful textual features (novelty, specificity, clout) rather than superficial shortcuts.

Our experiments span six datasets across two scientific domains and three publication years. Impact-DPO achieves the highest pairwise accuracy, reaching up to 83% and outperforming strong

embedding and prompting baselines by large margins. Compared with the main Sentence-T5-XL binary-classification baseline, Impact-DPO delivers an average improvement of about 30 percentage points, and supplementary same-backbone Qwen 2.5–7B BC reruns remain near chance, reinforcing that the gain comes from preference alignment rather than backbone capacity alone. Impact-DPO also improves by over 15 points relative to zero-shot prompting. These findings show that temporal prompting and preference alignment jointly enable large language models to reason effectively about dynamic influence, offering a scalable and interpretable solution for predicting future impact in evolving scholarly networks.

2 Related Work

Understanding and forecasting scientific impact have long been pursued through diverse methodological traditions in information retrieval, natural language processing, and network science. Prior work can be grouped into three broad strands, namely: text-based citation prediction, large-language-model approaches that rely on prompting or scalar scoring, and recent advances in preference-aligned learning that provide new tools for relative evaluation. Our work connects these directions through a text-only preference-learning framework that incorporates temporal reasoning.

2.1 Predicting Future Citations from Text

Early approaches estimated scientific impact directly from textual signals by encoding a paper's title or abstract into a fixed vector representation and training a supervised regressor or classifier against future citation targets [69]. The advent of transformer-based document encoders significantly advanced this line of work. SPECTER [9] introduced a domain-specific transformer trained on citation-informed objectives, demonstrating that contextualized embeddings can approximate relational information between papers and enable downstream impact prediction via simple linear models. Building on this paradigm, the SciRepEval benchmark [52] standardized the evaluation of representation learning for scientific documents, and subsequent SPECTER variants (often called SPECTER2 [52]) have become widely adopted off-the-shelf encoders for bibliometric prediction tasks.

We adopt this embedding to regressor paradigm as one of our baselines, using frozen SPECTER2 embeddings with a LinearSVR model whose scalar predictions are compared pairwise within each test set. Beyond SPECTER-based embeddings, other sentence-level encoders such as Sentence-T5 [37] provide higher-capacity representations of textual semantics. We include a Sentence-T5-XL binary-classification variant (BC) as our main discriminative baseline, and later report supplementary same-backbone Qwen 2.5–7B BC reruns to test whether larger model capacity alone can explain the gap to preference alignment. These baselines collectively reflect the state of text-centric bibliometric prediction before the advent of alignment-tuned LLMs.

2.2 LLM-based Impact Prediction and Prompting

Large Language Models (LLMs) have recently been applied to bibliometric forecasting by prompting them to assign scalar "impact" or "quality" scores from paper metadata such as title and abstract. De Winter [11] explored prompting ChatGPT to rate manuscripts along multiple qualitative dimensions and found moderate correlations with future citations but substantial variance across prompts and domains. We reproduce two prompting baselines inspired by this approach, i.e., a rating-style prompt ("Original") that aggregates multiple dimensions into a single impact value, and a simplified prompt ("Improved") that elicits a single normalized score. The resulting scalar outputs are converted into pairwise decisions for comparability with our setting.

A closer parallel to our work is the NAIP model ("newborn article impact prediction") [74], which finetunes an LLM to predict normalized impact directly from title and abstract. We treat NAIP as a released black-box model and evaluate it *as-is* on our standardized pairwise test sets. Although NAIP represents one of the first attempts to specialize LLMs for citation prediction, it remains anchored to scalar regression objectives. In contrast, our model operates purely on preference pairs and learns a comparative decision function that better mirrors how influence is evaluated. Empirically, our preference-aligned model trained on abstracts alone outperforms both prompting baselines and NAIP, hinting at the limitations of scalar calibration and zero-shot prompting for impact prediction.

2.3 Preference Learning and Direct Preference Optimization

Preference learning offers a principled alternative to regression or classification when the target signal is inherently comparative. Direct Preference Optimization [46] optimizes a policy to choose preferred outputs over dispreferred ones relative to a reference model, avoiding explicit reward modeling or reinforcement learning rollouts. While preference optimization has primarily been used for instruction following and model alignment, it has not previously been applied to forecasting *scientific impact* from text. Our problem is intrinsically relative, i.e., determining which of two referenced papers will accrue more citations by a given horizon, and therefore lends itself naturally to a pairwise preference-learning formulation. We use direct preference optimization to train a text-only policy over abstract inputs with minimal temporal metadata (publication years and horizon). This design departs from *(i)* scalar regression targets employed by embedding-based baselines and NAIP, and *(ii)* zero-shot prompting that lacks task-aligned optimization.

Existing approaches can thus be organized along two methodological axes: **(1)** whether they operate on absolute or relative supervision, and **(2)** whether temporal dynamics are explicitly represented. Embedding regressors such as SPECTER2+SVR are strong, simple, and label-efficient, but they predict absolute citation counts and rely on frozen encoders not optimized for comparative reasoning. Prompt-only LLMs are flexible but highly sensitive to prompt design and calibration, leading to unstable predictions without fine-tuning [11]. Released LLM scorers such as NAIP are fine-tuned for scalar regression from *title+abstract* inputs but remain untested under comparative evaluation. Impact-DPO distinguishes itself in three key respects. First, it is trained directly on relative supervision, aligning the model with the pairwise decision structure of the task. Second, it introduces temporal cues into text-only prompts, allowing temporal reasoning without explicit graph propagation. Third, it isolates the contribution of preference alignment by maintaining strict comparability with prior text and embedding baselines. Together, these distinctions allow us to evaluate the unique value of preference alignment for text-based impact prediction while situating our work within the broader evolution of LLM-based bibliometric modeling.

3 Formal Problem Definition

The objective of this work is to model and predict the relative future influence of scholarly documents based on their textual and temporal characteristics within a dynamic citation network. Influence is not treated as an absolute scalar quantity but as a comparative property that emerges over time as documents interact through citations. The goal is therefore to determine, for a given context of citing and cited works, which of two candidate papers will accumulate greater influence at a future horizon. This formulation captures the inherently relational nature of scientific attention and provides a natural foundation for preference-based learning.

We consider a temporal text-attributed graph (TAG) $G(t) = (V_t, E_t)$, where each node $v \in V_t$ corresponds to a document (e.g., a research paper) and each directed edge $(v_i, v_j) \in E_t$ represents a citation from v_i to v_j observed up to time t. Each node is associated with an attribute vector A_v

containing textual and temporal information such as the paper's abstract, title, and publication year. The graph evolves over time as new papers and citations are introduced, resulting in changing connectivity and influence patterns.

The task of *pairwise future-citation prediction* is defined as follows. For a given citing paper v observed at time t, let v' and v'' be two distinct papers that are both cited by v. The objective is to predict which of these two cited papers will exhibit a higher cumulative measure of influence by a future time $T > t$. Influence is quantified through a cumulative function $C(v, v', T)$, which aggregates the total citations received by paper v' from its publication up to the prediction horizon T.

Formally, we seek to learn a function f_θ parameterized by model weights θ, such that

$$f_\theta(v, v', v'', t, T) = \begin{cases} 1, & \text{if } C(v, v', T) > C(v, v'', T) \\ 0, & \text{otherwise.} \end{cases}$$

The model receives as input the textual and temporal representations of v, v', and v'', typically in the form of abstract-level embeddings or prompt-conditioned sequences. The training objective is to learn parameters θ that produce accurate pairwise comparisons across all observed citation contexts.

This formulation generalizes to any temporal graph where influence can be measured as an accumulative attribute of nodes, such as retweet counts in social networks or citation frequencies in patent databases. In the specific case of academic citation networks, the temporal text-attributed graph $G(t)$ captures:

- **Nodes.** Each node $v \in V_t$ corresponds to a scientific paper. The attribute A_v includes its textual content (title and abstract) and metadata (e.g., publication year).
- **Edges.** Each directed edge $(v_i, v_j) \in E_t$ represents a citation from paper v_i to paper v_j occurring at time t. New edges appear as new papers cite existing ones, and thus the graph structure evolves dynamically.

For each citing paper v, we construct training instances consisting of its cited pairs (v', v''). Each instance is associated with a label derived from the relative citation accumulation of the two cited papers by horizon T. The dataset thus comprises a set of pairwise preference observations

$$\mathcal{D} = \{(v, v', v'', y) \mid y = \mathbb{I}[C(v, v', T) > C(v, v'', T)]\},$$

where $\mathbb{I}[\cdot]$ is the indicator function. Each instance defines a local comparative judgment of future influence conditioned on the citing context v.

This problem formulation captures several key characteristics of scientific impact prediction. First, it emphasizes relative comparison rather than absolute estimation, reducing sensitivity to domain-level citation inflation and temporal drift. Second, it preserves the natural relational structure of scholarly communication by conditioning predictions on shared citing papers. Finally, it aligns closely with preference-based learning paradigms that optimize directly for relative correctness, providing a more faithful representation of how influence emerges and is recognized in dynamic scientific networks.

4 Proposed Approach

Current approaches capture either the semantic or structural aspects of citation behavior but fall short of modeling how influence evolves within dynamic, text-attributed networks. Text-only methods excel at capturing semantic similarity but fail to incorporate the temporal context in which citations accumulate. Graph-based models, in turn, encode network dependencies but are often computationally intensive and poorly suited for long-horizon inference [6, 27]. These limitations

underscore the need for an approach that jointly considers temporal, structural, and textual signals within a unified modeling framework.

We address this challenge through a method that leverages *preference alignment with temporal reasoning* using large language models. Impact-DPO is designed to align the model's predictive behavior with observed future influence patterns while maintaining interpretability and scalability. The formulation rests on three interconnected principles, namely **(1)** preference learning through direct optimization, **(2)** temporal reasoning via structured prompting, and **(3)** the integration of textual and structural cues in a text-only inference setting.

The first component of our method is the reformulation of the task as a *preference-learning problem* rather than as regression or classification. Instead of predicting absolute citation counts, the model learns to compare pairs of papers and determine which will accrue greater influence within a specified future time period. This reframing aligns directly with the pairwise formalism introduced in Section 3 and builds on advances in preference-based optimization that have proven effective for aligning large language models with human or task-specific judgments [46, 65]. By focusing on relative differences, the model avoids issues of scale sensitivity and domain heterogeneity that often affect citation regression models.

The second design element introduces *temporal reasoning through prompting*. Citation networks evolve continuously as new papers and edges are added, and the relevance of textual signals changes over time [39, 63]. To capture this temporal evolution, we construct input prompts that encode temporal context directly into the language model's input space. Each prompt provides the textual attributes of a citing paper v and its two cited papers v' and v'', augmented by lightweight temporal metadata such as publication years and the prediction horizon T. This encoding enables the model to reason about recency, order, and temporal intervals without modifying its underlying architecture. It also aligns with recent work showing that language models can effectively handle structured temporal context through carefully designed prompting strategies [43, 68].

Importantly, our method integrates textual semantics and temporal structure while maintaining a text-only inference model. The citation graph is not used for message passing or adjacency encoding during model inference; instead, it plays a crucial role during data construction and labeling. Specifically, the graph is employed to *(i)* generate training and evaluation pairs, i.e., two cited papers connected through a common citing node, and *(ii)* determine which of the two is preferred based on observed cumulative citation counts in the time period T. This design maintains the relational grounding of graph-based models while ensuring that inference remains efficient, interpretable, and adaptable across domains. By disentangling representation learning from graph traversal, we leverage the generalization capacity of large language models without incurring the overhead of structural computation [52, 63].

The temporal prompting mechanism operationalizes this integration by embedding both textual and temporal features into a unified representation. For each citing paper v, the prompt includes the abstracts of v' and v'', their respective publication times, and the prediction time period T. These inputs allow the model to reason about the temporal spacing between publication and evaluation times and to infer how textual similarity interacts with recency and topic trends. This explicit encoding of time and text enhances the model's ability to identify evolving influence patterns, such as the rapid rise of newly relevant topics or the long-term persistence of foundational works [39, 63]. Through this design, temporal awareness is achieved not by architectural modification but through structured conditioning that exploits the inherent sequence reasoning capacity of transformer-based models.

Formally, we model the task as maximizing the conditional probability:

$$p(C(v', T) > C(v'', T) \mid v),$$

where $C(v', T)$ and $C(v'', T)$ denote the future cumulative citations of the two papers v' and v'', respectively, and v represents the common citing paper that anchors the comparison. To align the model with this objective, we employ the *direct preference optimization* framework [46], which enables direct optimization over preference pairs without explicit reward modeling. The loss is given by

$$\mathcal{L}(\pi_{\text{old}}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_u, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\text{old}}(y_u \mid x)}{\pi_{\text{ref}}(y_u \mid x)} - \beta \log \frac{\pi_{\text{old}}(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right], \tag{1}$$

where π_{old} is the model policy, π_{ref} is a reference policy (typically the pretrained base model), y_u and y_l represent the preferred and non-preferred outputs, $\sigma(\cdot)$ is the sigmoid function, and β controls the relative weighting of the preference margin. The training data distribution $\mathcal{D}$ consists of triplets (x, y_u, y_l), where x is the textual and temporal input prompt corresponding to a citing context v, and y_u and y_l correspond to the two cited papers, labeled according to observed citation outcomes by horizon T.

In our setting, the preferred output y_u corresponds to the paper v' that has accrued more citations than v'' by time T, while y_l represents the less influential counterpart. Optimizing Equation 1 trains the model to assign higher conditional likelihood to the preferred text sequence, thereby maximizing $p(C(v', T) > C(v'', T) \mid v)$. This objective directly reflects the pairwise decision structure of the problem and promotes sensitivity to subtle textual and temporal cues that differentiate future influence trajectories. The preference-based learning formulation thus aligns naturally with the dynamic and relational characteristics of citation networks, encouraging the model to internalize both semantic relevance and temporal momentum.

From a theoretical standpoint, our formulation can be justified through the correspondence between pairwise preference learning and margin-based risk minimization in non-stationary temporal environments [2, 10, 42] such as evolving citation networks. Let each paper be represented by a latent embedding $\phi(v; \theta) \in \mathbb{R}^d$ that encodes both its textual and temporal properties through the parameterization of the large language model. The objective is to learn a scoring function $s(v; \theta) = \langle w, \phi(v; \theta) \rangle$ whose induced order relation $s(v') > s(v'')$ approximates the true but unobservable ordering of scholarly influence, defined by the cumulative citation trajectories $C(v', T) > C(v'', T)$. Under standard assumptions of pairwise consistency and transitivity in ranking functions [46], minimizing the objective (Equation 1) is equivalent to minimizing a convex upper bound of the expected pairwise ranking risk:

$$\mathcal{R}(\theta) = \mathbb{E}_{(v, v', v'')} \left[\ell\big(s(v'; \theta) - s(v''; \theta),\ \text{sign}(C(v', T) - C(v'', T))\big) \right],$$

where $\ell(\cdot)$ is a monotonic surrogate loss such as the logistic or hinge loss. In this formulation, the model learns to approximate the latent "citation potential" of papers through their textual and temporal representations, effectively capturing how meaning and timing interact to shape influence within the scientific corpus. The introduction of temporal prompting acts as a conditioning mechanism that renders $\phi(v; \theta)$ sensitive to both semantic and chronological cues, transforming a non-stationary citation prediction problem into a conditionally stationary one with respect to time-conditioned distributions $P_t(v)$. This ensures that embeddings for papers published in different years remain comparable while reflecting domain-specific citation dynamics such as recency bias, delayed recognition, or field maturation [24, 38, 63]. Direct preference optimization, in turn, enforces a maximum-margin alignment between predicted and empirical citation orderings, thereby aligning model predictions with the observed evolution of paper influence. Collectively, these mechanisms enable the model to approximate an optimal preference policy π^* that minimizes expected citation misordering across temporal contexts, providing theoretical justification for its ability to generalize under the evolving citation patterns over years.

5 Experimental Setup

This section outlines the experimental framework designed to evaluate Impact-DPO for forecasting citation influence. We first articulate the research questions guiding our analysis, then describe the datasets, preprocessing pipeline, and temporal pair construction. Finally, we detail the baseline systems and their operational definitions to ensure comparability across methods.

5.1 Research Questions

Our experiments are structured around four research questions (RQs), each addressing a distinct evaluative dimension:

RQ1 (Predictive performance). Does Impact-DPO outperform strong text-embedding, prompting-only, and released-scorer baselines on pairwise future-influence prediction across multiple scientific domains and temporal snapshots?

RQ2 (Robustness). How consistent is Impact-DPO's performance across temporal splits, domain-specific citation dynamics, and backbone LLM architectures?

RQ3 (Ablation). What is the individual contribution of the preference-alignment objective relative to (a) standard binary-classification finetuning on the same backbone and (b) zero-shot prompting without task-specific supervision?

RQ4 (Sensitivity & generative alignment). How do hyperparameters, particularly the preference scaling factor β, influence learning stability, and to what extent does the aligned model generate text that semantically resembles high-impact papers?

These questions collectively assess the effectiveness, robustness, and generalizability of preference-aligned LLMs for temporal influence modeling.

5.2 Datasets

We evaluate all models on two large-scale temporal text-attributed graphs derived from two bibliometric domains, namely *Computer Science (CS)* and *Physics*. Each domain is processed independently using an identical filtering, normalization, and temporal pairing pipeline to ensure comparability and prevent cross-domain leakage.

Data preprocessing. We begin by filtering the raw corpus to include only papers that have received at least one citation, ensuring that each node in the citation graph exhibits measurable scholarly influence. We exclude papers whose titles or abstracts include terms indicative of survey articles, software repositories, or COVID-19–related studies to avoid confounding factors such as extraordinary citation bursts or review effects [3, 4]. Text cleaning is applied to standardize abstracts and titles, and citation metadata are parsed using the `anystyle` toolkit[1], which employs conditional random fields for bibliographic normalization. This step ensures consistent linking between citing and cited papers.

Each paper node v in the graph is thus represented by its title, abstract, publication year, and cumulative citation count. Edges denote directed citation relationships, producing a temporal citation graph $G(t) = (V_t, E_t)$ where new nodes and edges are progressively added as publication years advance.

Temporal preference construction. To capture comparative influence dynamics, we transform each domain-specific graph into a temporally ordered set of preference pairs. For a given citing paper v and its referenced papers v' and v'', we construct a training instance (v, v', v'', y), where $y = 1$ if $C(v', T) > C(v'', T)$ and $y = 0$ otherwise, for a prediction horizon $T > t_v$. To mitigate citation skew, we retain only pairs where the citation gap exceeds the median difference for the publication

[1] https://github.com/inukshuk/anystyle

Table 1. Dataset statistics for 2018–2020. Split (%) is the row's share of the domain–year total.

Domain	Year	Split	Nodes	Edges	Pairs	Split (%)
Computer Science	2018	Overall	15 250	54 864	27 435	100.00
	2018	Train	7382	34 293	17 147	62.50
	2018	Test	7984	20 571	10 288	37.50
	2019	Overall	15 250	54 864	27 435	100.00
	2019	Train	13 177	45 168	22 586	82.33
	2019	Test	4546	9696	4849	17.67
	2020	Overall	15 250	54 864	27 435	100.00
	2020	Train	14 593	51 058	25 532	93.06
	2020	Test	1947	3806	1903	6.94
Physics	2018	Overall	52 518	175 468	87 736	100.00
	2018	Train	46 796	145 052	72 528	82.67
	2018	Test	13 461	30 416	15 208	17.33
	2019	Overall	52 518	175 468	87 736	100.00
	2019	Train	48 323	155 138	77 571	88.41
	2019	Test	10 040	20 330	10 165	11.59
	2020	Overall	52 518	175 468	87 736	100.00
	2020	Train	49 609	163 464	81 734	93.16
	2020	Test	6826	12 004	6002	6.84

year. Concretely, for each publication year we compute the median of the absolute citation-count differences $|C(v', T) - C(v'', T)|$ across all candidate pairs in that year; a pair is kept only if its gap strictly exceeds this year-specific median, ensuring that the training signal focuses on pairs with clearly distinguishable influence trajectories rather than near-ties (Figure 1). Temporal consistency is maintained by enforcing a strict chronological split between training and testing subsets, namely models are trained on papers published before a cutoff year and evaluated on papers published after that cutoff. Separate datasets are constructed for 2018, 2019, and 2020, each providing a distinct temporal snapshot of scholarly influence. Cumulative citation counts $C(v, T)$ are measured at the end of the unarXive corpus's temporal coverage ($T = 2022$), yielding observation windows of approximately 2–4 years for the 2018–2020 cohorts. Because the 2020 cohort has only a ∼2-year window, some papers may not yet have reached their citation peak; results for that split should therefore be interpreted as reflecting medium-term rather than long-term influence, and extending the horizon as the corpus grows is a natural direction for future work.

Co-citation comparability and topic controls. A key advantage of our pair-construction strategy is that both candidates in each pair share a common citing paper v. Because v explicitly references both v' and v'', the two candidates are *co-cited*, a relationship long recognized as a strong indicator of topical proximity [35, 54]. Co-citation pairing therefore provides an implicit topic control, i.e., candidates compete within the same intellectual niche rather than across unrelated fields, reducing the confounding effect of cross-topic citation-rate differences. To quantify this, we computed SPECTER2 cosine similarities for all test pairs. Within-domain embeddings are inherently tightly clustered, yet co-cited pairs are significantly more similar than size-matched random pairs drawn

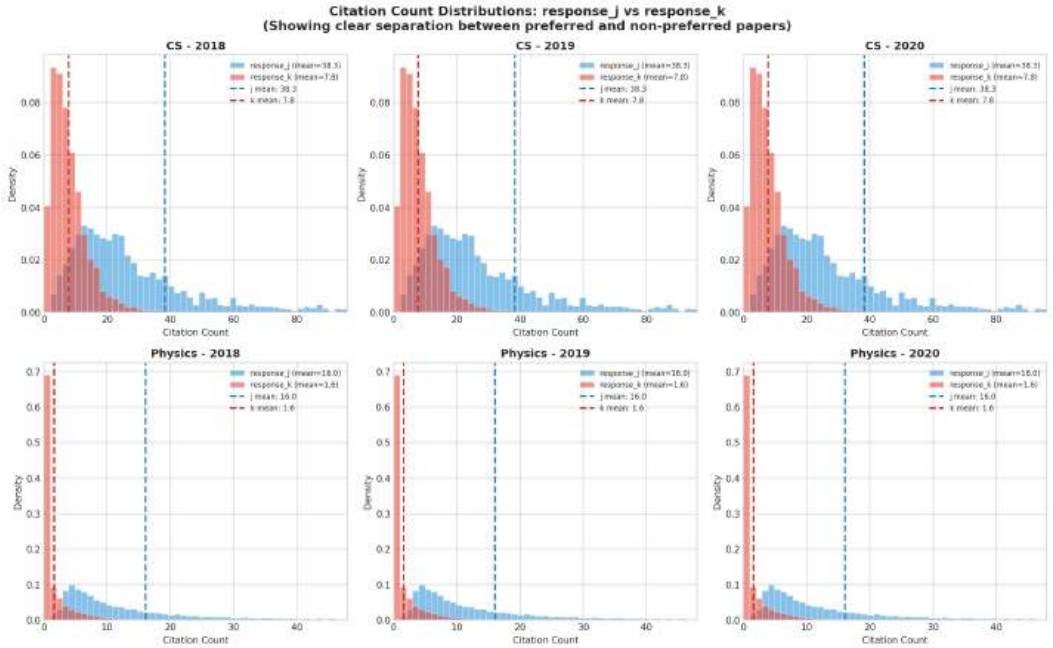

Fig. 1. Citation-count distributions for preferred (response_j, blue) and non-preferred (response_k, red) papers after median-gap filtering, shown per domain and year. In CS the preferred-paper mean is ≈38 citations versus ≈8 for non-preferred; in Physics the corresponding values are ≈16 versus ≈1.6. The clear separation confirms that the median-based filtering retains pairs with meaningfully distinct influence trajectories.

from the same domain: mean cosine similarity of 0.986 ± 0.010 vs. 0.978 ± 0.014 in CS and 0.984 ± 0.011 vs. 0.974 ± 0.015 in Physics (two-sample t-test, $t > 40$, $p < 0.001$ in all year splits). The absolute difference is modest because within-domain embeddings are already tightly clustered; nevertheless, the effect is consistent across all six domain-year conditions, indicating that co-citation pairing provides an additional, reliable layer of topical proximity beyond shared domain membership alone. This assumption is not universal where some co-cited pairs combine, for example, a broad methods paper with a domain-specific application paper that are cited together for different rhetorical roles. We therefore treat co-citation as a strong but noisy proxy for comparability rather than a substitute for explicit within-topic stratification. We further restrict the corpus to single-domain subsets (Computer Science or Physics) and filter out surveys, software papers, and COVID-19 studies, which together limit heterogeneity in citation dynamics. While explicit topic-model controls (e.g., stratifying pairs by LDA cluster) could refine comparability further, the co-citation constraint already ensures that most pairs evaluated by the model reflect meaningful within-niche comparisons rather than arbitrary cross-field rankings.

The resulting dataset is thus both temporally grounded and structurally balanced, reflecting realistic citation dynamics across time. Table 1 reports detailed statistics, including node, edge, and pair counts for each domain-year combination. As shown, Computer Science graphs are smaller but more topically heterogeneous, while Physics graphs exhibit higher density and a more pronounced long-tail citation distribution. Figure 2 visualizes this heavy-tailed degree centrality pattern, consistent with the power-law structure of academic citation networks [19, 44].

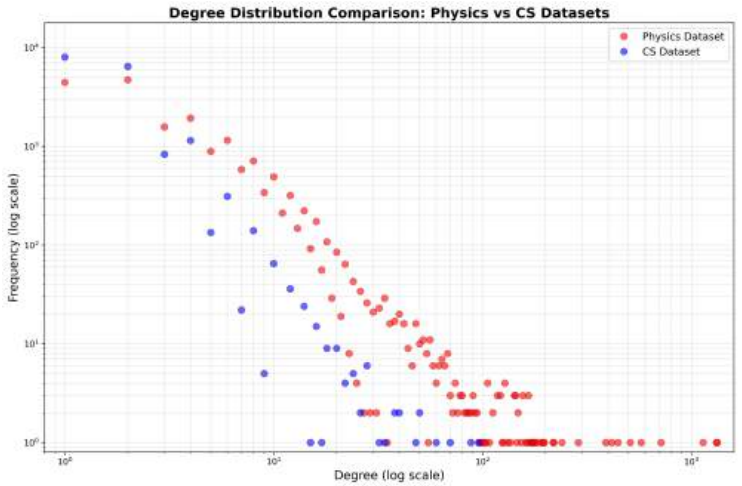

Fig. 2. Degree Distribution of the Dataset Graph Regardless of the Cutoff Threshold of both CS and Physics Datasets.

Dataset integrity. To prevent data leakage between temporal partitions, the citing papers that form the basis of pair construction are strictly disjoint across train and test splits. Each method is evaluated on the same fixed set of test pairs, ensuring that observed differences in performance are attributable solely to modeling strategies rather than data partitioning. When models output continuous influence scores rather than direct pairwise labels, scores are converted to binary comparisons to maintain evaluative consistency. All models are evaluated on identical pairwise test sets for each domain and year. Pairwise accuracy is used as the primary metric, capturing the proportion of correctly ordered influence pairs. For methods that produce scalar impact scores, decisions are derived by simple comparison $\hat{y} = \mathbb{I}[s(v') > s(v'')]$. Hyperparameters are tuned on a held-out validation subset of the 2018 dataset and reused for subsequent years to ensure cross-temporal consistency. To quantify sampling uncertainty, we report 95% bootstrap confidence intervals (1 000 resamples of the test pairs) for all headline accuracy figures; where intervals are omitted from tables for space, they are available in the replication package.

5.3 Baseline Methods

To rigorously evaluate Impact-DPO, we compare it against four classes of baselines that collectively represent the main paradigms of scholarly influence prediction. These baselines are chosen to isolate the contribution of preference alignment, assess the role of temporal prompting, and benchmark against established text and embedding pipelines: **(1)** *Binary finetuning variant (BC)* [21, 47, 64] serves as the most direct comparator to our method. Our primary BC instantiation uses Sentence-T5-XL as a strong discriminative sequence classifier over paired abstracts, and Section 6.3 additionally reports supplementary same-backbone Qwen 2.5–7B BC reruns to separate objective effects from backbone size. This setup allows us to quantify the benefit of Impact-DPO by contrasting it with conventional supervised finetuning, which optimizes for absolute label correctness rather than relative preference satisfaction. **(2)** *Embedding-based regression (SPECTER2+SVR)* represents the established embedding–regressor paradigm widely used in bibliometric prediction pipelines [52]. In this configuration, frozen document embeddings are generated using SPECTER2 [52], a transformer trained on citation-aware objectives, and a linear Support Vector Regressor (SVR) is fit to predict

continuous citation scores. Pairwise decisions are obtained by comparing the predicted scalar values within each pair. This baseline is strong, label-efficient, and interpretable but inherently limited by its reliance on static embeddings that cannot adapt to evolving temporal or contextual nuances in citation patterns. **(3)** *Released impact scorer (NAIP)* offers a domain-specific, fine-tuned LLM that predicts normalized impact scores from paper titles and abstracts [74]. We evaluate this model *as-is*, without retraining, to assess the zero-shot generalization capacity of an existing LLM trained for scalar citation forecasting. This baseline provides an important point of comparison for understanding how specialized impact scorers perform relative to our preference-aligned framework when evaluated under identical pairwise conditions. **(4)** *Prompting-only predictors* extend the line of research exploring in-context citation forecasting with LLMs [11]. We consider two representative configurations: an *original* prompt that elicits multi-dimensional ratings (e.g., novelty, rigor, clarity) subsequently aggregated into a composite impact score, and an *improved* single-score prompt that directly requests a normalized impact estimate in the range $[0, 1]$. These prompting baselines test whether an unaligned LLM can reason about scientific influence purely through context engineering, thereby serving as a control for finetuning effects.

We provide all prompt templates and model specifications in the Appendix of this paper.

6 Results and Findings

6.1 Comparison to Text and Embedding Baselines (RQ1: Predictive Performance)

The first stage of our evaluation investigates whether our preference-aligned model can more accurately forecast relative scholarly influence than traditional text-embedding regressors and prompting-only predictors. All models are assessed on identical pairwise test sets constructed for each domain–year combination (Computer Science and Physics; 2018–2020). When a model outputs scalar citation scores, such as SPECTER2+SVR, NAIP, or prompted LLMs, these values are converted to pairwise decisions by directly comparing the two within each pair. Each method operates under a uniform text-only input regime, where the title and abstract serve as the primary textual descriptors and publication year provides the temporal cue. The citation graph is used exclusively for forming training pairs and defining future citation-based preferences, ensuring comparability across models.

Across both domains, Impact-DPO exhibits the strongest and most consistent performance, achieving the highest accuracy across all temporal snapshots. Its advantage holds under diverse conditions of citation density and domain heterogeneity. Embedding-based methods, typified by SPECTER2+SVR, remain competitive but plateau in their predictive capacity, likely due to the static and non-comparative nature of their learned representations. The released NAIP model, though specifically fine-tuned for impact prediction, underperforms in our pairwise evaluation setting, suggesting that absolute citation forecasting does not directly translate into accurate comparative ranking. Prompt-only and zero-shot models display even weaker performance, underscoring the insufficiency of unsupervised language priors for learning temporal influence orderings.

The performance patterns observed are consistent with theoretical expectations from prior work on document representation and influence modeling. Static embedding models such as SPECTER2 encode fixed relational priors derived from citation co-occurrence at training time [52], which limits their ability to generalize to newly formed citation relationships or temporally drifting discourse. Prompting-only LLMs, although contextually flexible, lack the explicit supervision necessary to learn preference-consistent orderings and thus default to linguistic heuristics that correlate weakly with future impact [74]. The Impact-DPO training objective directly addresses these limitations whereby optimizing for relational correctness rather than absolute prediction error, the model internalizes the comparative dependencies that drive citation accumulation. Temporal prompting

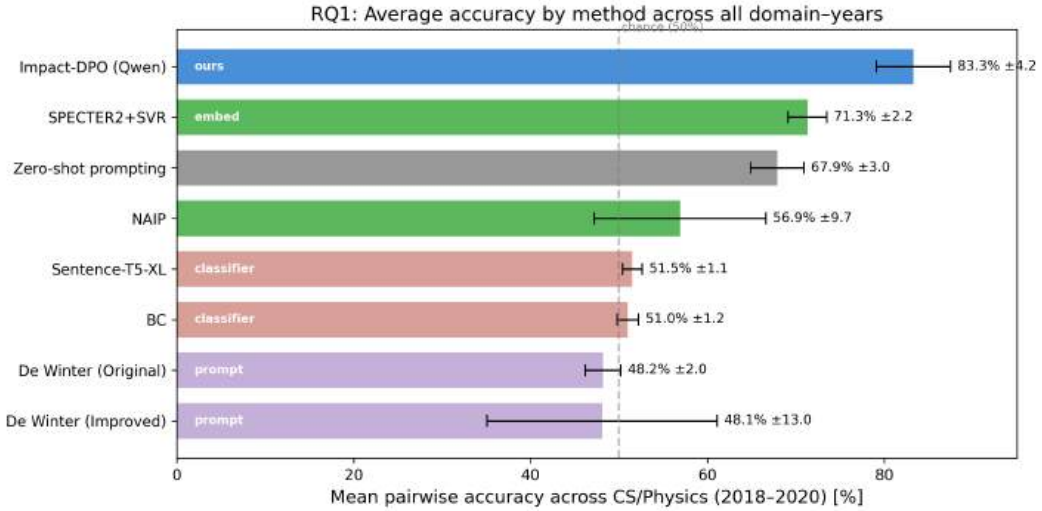

Fig. 3. **RQ1: Average accuracy by method across all domain–years.** Bars show mean pairwise accuracy pooled over CS/Physics (2018–2020); error bars indicate 95% bootstrap confidence intervals. The dashed line marks chance performance (50%). Colors group methods by family (ours / embedding / prompt / released / other).

further grounds this reasoning by situating each prediction within its publication context, effectively conditioning the model's representation on the evolving temporal distribution of attention.

Overall, RQ1 findings indicate that preference alignment equips large language models with the capacity to infer relative scholarly influence directly from textual and temporal cues. The consistent gains across domains show that Impact-DPO is an effective and generalizable framework for forecasting future citation orderings.

6.2 Performance Consistency (RQ2a: Temporal Robustness)

The second part of our analysis examines whether model performance remains consistent across time and domain, focusing on the stability of both absolute accuracies and relative method rankings. Each model is evaluated on three consecutive temporal snapshots (2018–2020) within the Computer Science and Physics datasets, which differ in both citation density and topic evolution. To quantify stability, we summarize mean accuracy, inter-year variation, and trend direction (percentage points per year) for each method, providing a measure of robustness under temporal and domain shifts.

Across all years and domains, the relative ordering of models remains unchanged, i.e., Impact-DPO outperforms embedding-based regressors, followed by released impact scorers and prompting-only baselines, with binary classifiers and sentence encoders performing near random. In Computer Science, Impact-DPO consistently leads across all three years, showing only a mild decline from 2020 to 2018, a pattern plausibly explained by increasing pair difficulty in earlier years where citation gaps are smaller and influence trajectories less distinct. The same hierarchical ordering holds in Physics, where Impact-DPO maintains a stable margin over SPECTER2+SVR and NAIP, suggesting that the advantages conferred by preference alignment are not domain-specific.

The stability of the performance ordering across temporal splits is theoretically consistent with expectations from non-stationary learning and representation dynamics. Temporal shifts in citation networks alter the marginal distributions of topics and publication venues but not the fundamental

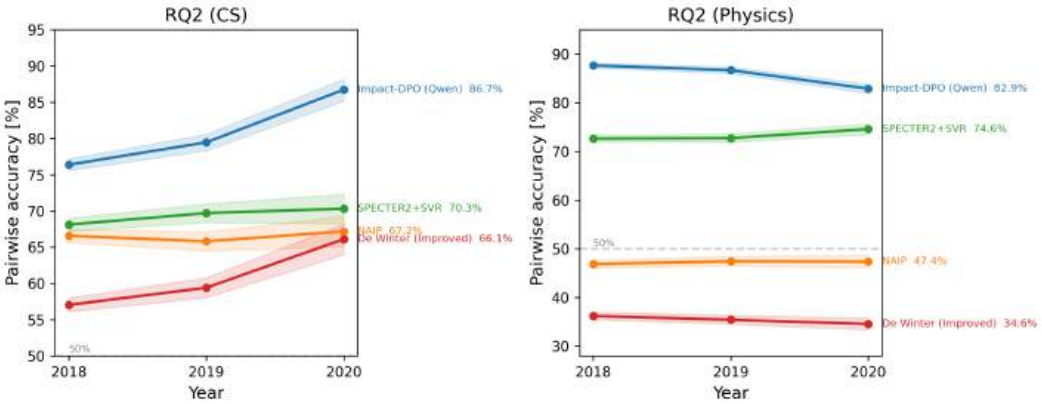

Fig. 4. **RQ2: Accuracy by year for representative methods.** Impact-DPO remains consistently superior across temporal snapshots for both CS and Physics Datasets across various cutoffs, while SPECTER2+SVR exhibits limited change and prompting-only models trail with smaller gains. Shaded bands show 95% bootstrap confidence intervals.

relational property that high-impact papers attract disproportionately more citations over time [63]. Since Impact-DPO directly optimizes over these relative preferences, it learns representations that are invariant to such marginal temporal drift. Embedding-based regressors, by contrast, depend on fixed relational priors established during pretraining, which limits adaptability to new temporal distributions. Similarly, prompting-only LLMs lack an explicit temporal conditioning mechanism, making them more susceptible to contextual shifts and topic drift.

Overall, the results indicate that preference alignment with temporally explicit prompting yields consistent and domain-robust performance across multiple years of evaluation. The observed stability in both absolute and relative rankings supports the view that optimizing over empirical citation orderings confers temporal generalization and allows the model to preserve predictive accuracy despite evolving citation behaviors.

6.3 Binary Classification versus Preference Alignment (RQ3a: Objective Ablation)

This analysis compares Impact-DPO against a strong binary-classification baseline on identical pairwise test sets in the Computer Science domain. The main BC baseline uses Sentence-T5-XL rather than a 7B generative backbone, so Table 2 should be read as the primary objective comparison against a competitive discriminative classifier. To directly address the remaining capacity question, we also ran supplementary same-backbone BC experiments with Qwen 2.5–7B; those results are summarized after Table 2.

Replacing BC with Impact-DPO produces large and consistent improvements across all temporal splits. Impact-DPO achieves over 20 percentage points higher accuracy on average than its BC counterpart, demonstrating that aligning the optimization objective with the pairwise evaluation criterion substantially enhances learning efficacy. These trends persist across all years as well.

The theoretical basis for these improvements lies in the fundamental difference between how the two objectives encode relational structure. Binary classification forces the model to learn an absolute decision boundary in the concatenated input space, implicitly assuming that the underlying score distributions for influential and non-influential papers are separable in feature space [10]. This assumption rarely holds in citation forecasting, where influence is continuous and context-dependent. Impact-DPO, by contrast, directly optimizes for the relative ordering that

Table 2. **RQ3 (CS):** Accuracy (%) for *BC* vs *Impact-DPO*. Improvements reported in percentage points on the CS dataset, the same results hold on the Physics dataset (*Impact-DPO − BC*).

Year	BC	Impact-DPO	Improvement (pp)
2018	51.72	76.38	+24.66
2019	51.67	79.46	+27.79
2020	52.75	86.73	+33.98
Mean	52.05	80.86	**+28.81**

defines the evaluation metric. It models the conditional probability that one paper will accrue more citations than another, effectively learning a monotonic mapping from latent semantic–temporal representations to influence orderings. By integrating temporal prompts, Impact-DPO further captures how comparative salience shifts over time, such as the transient rise of emerging topics or the delayed recognition of foundational works, which a fixed binary boundary cannot express.

In summary, preference alignment yields at least a 20-point improvement over binary finetuning. These findings confirm that Impact-DPO not only enhances optimization alignment with the evaluation objective but also enables the model to exploit nuanced temporal and semantic cues that conventional classification losses fail to capture. To directly test whether this large gap could be explained by backbone capacity rather than the training objective, we additionally fine-tuned Qwen 2.5–7B with a BC objective under a matched sequence-classification setup. Even with the same Qwen backbone, BC remained near chance across all six splits: CS 2018 55.64% [54.95, 56.34], CS 2019 58.08% [57.09, 59.02], CS 2020 51.90% [50.29, 53.55], Physics 2018 51.44% [50.69, 52.14], Physics 2019 53.65% [52.73, 54.57], and Physics 2020 50.02% [48.56, 51.50] (95% bootstrap CIs). Physics 2020 is statistically indistinguishable from random guessing because its interval crosses 50%. These supplementary results materially weaken the capacity-confound explanation in that upgrading BC to the same Qwen backbone still does not approach Impact-DPO, so the dominant difference remains the preference-alignment objective.

6.4 Comparison with Zero-Shot Setting (RQ3b: Zero-Shot Ablation)

This analysis evaluates how task-specific fine-tuning proposed in this paper alters the capabilities of large language models compared to their zero-shot counterparts. Both settings use the same backbone and are evaluated on identical pairwise test sets across the Computer Science and Physics datasets from 2018–2020. While Impact-DPO is explicitly aligned to the temporal and relational structure of citation graphs, the zero-shot variant relies solely on pre-trained world knowledge without exposure to task-specific supervision.

Across all domains and years, Impact-DPO exhibits a substantial and consistent advantage over the zero-shot counterparts. The zero-shot configuration shows an accuracy of ranging from 64–70%, while Impact-DPO achieves between 76–88%. The per-split advantage spans from approximately +6 to +23 percentage points, averaging +15 pp across six test sets. These results demonstrate that pre-trained LLMs, though capable of rich language understanding, lack the inductive bias necessary to infer temporal dependencies and relative influence patterns that govern citation dynamics.

From a theoretical standpoint, this gap arises because zero-shot LLMs operate on static priors learned from general text corpora that lack temporally indexed relational supervision. Citation influence prediction, however, depends on modeling evolving dependencies between publication time, semantic novelty, and structural diffusion within the scholarly network [8, 55]. Without explicit alignment to these temporal cues, a zero-shot model cannot distinguish between contemporaneous

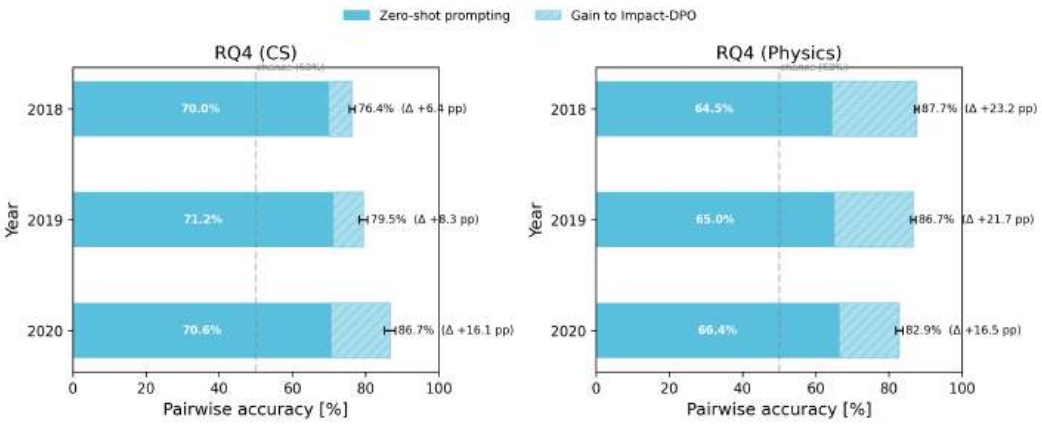

Fig. 5. **RQ3b: Impact-DPO vs. zero-shot (per split).** Side-by-side accuracies for each domain–year; whiskers show 95% bootstrap confidence intervals. The dashed line marks chance (50%). Impact-DPO consistently dominates zero-shot performance across all six splits.

and historically lagged influence patterns. Fine-tuning Impact-DPO addresses this by optimizing on observed pairwise outcomes, it teaches the model to encode relative temporal progression and cumulative advantage effects intrinsic to citation growth. In doing so, Impact-DPO reorients the model's latent space toward causal-temporal reasoning rather than static semantic similarity.

6.5 Impact of Choice of LLM (RQ2b: Architecture Robustness)

This analysis investigates whether the performance of preference-aligned models depends primarily on the choice of backbone language model or on the alignment objective itself. To isolate the effect of architecture, we compare two variants, namely LLaMA2–7B and Qwen2.5–7B, on identical Computer Science splits (2018–2020), while holding all training procedures, prompts, and data constant.

As shown by Figure 6 across all three years in the Computer Science dataset, Qwen2.5–7B consistently surpasses LLaMA2–7B by 3–14 percentage points, yielding a mean improvement of roughly +9.6 pp. Importantly, both LLMs, regardless of absolute scale, substantially outperform all baselines, including SPECTER2+SVR and NAIP, by wide margins. These results indicate that the majority of the performance gain arises from the preference-alignment mechanism and temporally explicit prompting, with model capacity influencing only the achievable upper bound rather than the direction of improvement.

6.6 Sensitivity to Preference Strength Parameter (β) (RQ4a: Hyperparameter Sensitivity)

This analysis examines how the temperature-like parameter β in our loss influences model behavior and accuracy across years. Figure 7 illustrates that smaller β values, particularly $\beta = 0.1$, consistently yield the highest performance across all temporal splits. In contrast, higher values such as $\beta = 0.8$ produce a systematic decline in accuracy, with the 2020 dataset showing a reduction from approximately 86.7% at $\beta = 0.1$ to 76.5% at $\beta = 0.8$. This trend shows the central role of β in mediating the trade-off between model flexibility and adherence to the reference distribution.

From a theoretical standpoint, β controls the degree of alignment regularization in Impact-DPO. Lower values down-weight the influence of the reference policy π_{ref}, allowing the fine-tuned model

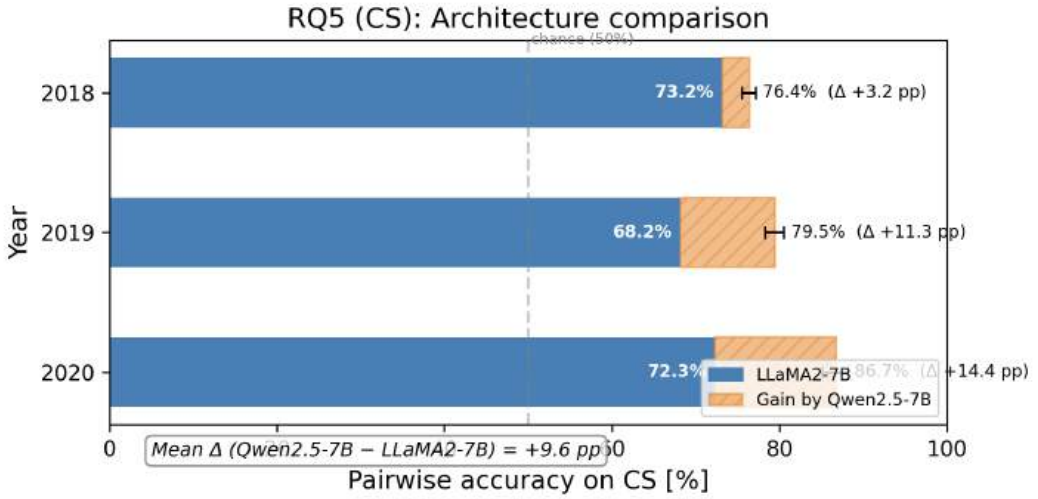

Fig. 6. **RQ2b (CS): Impact of the choice of LLM on the performance of Impact-DPO.** For each year, the solid bar shows *LLaMA2-7B* accuracy and the hatched extension shows the gain from *Qwen2.5-7B*; whiskers indicate 95% bootstrap confidence intervals. Qwen exceeds LLaMA2 by +**3.2** / +**11.3** / +**14.4 pp** in 2018/2019/2020 (mean +**9.6 pp**).

π_{old} to deviate more freely when the empirical data exhibit ambiguous or overlapping preferences. In contrast, larger β values impose stronger regularization, effectively constraining the model to remain closer to the reference distribution even when the task demands finer discrimination. In domains such as citation influence prediction, where the semantic and temporal signals that differentiate highly cited papers from moderately cited ones are subtle, this increased flexibility becomes essential for accurate learning.

The empirical pattern is consistent with what we term a *low-gap preference regime*, where differences between preferred and non-preferred instances are small relative to model uncertainty. Citation forecasting plausibly exemplifies such a regime: most pairs of papers share topical overlap and comparable early visibility, and their eventual citation divergence emerges gradually over time. Under these conditions, low β values permit the model to remain sensitive to subtle textual and temporal cues, thereby improving discrimination among influence trajectories with small initial differences. Conversely, high β values suppress this adaptive behavior by over-penalizing divergence from the reference model, leading to underfitting.

6.7 Ability to Generate Future Nodes (RQ4b: Generative Alignment)

This research question examines whether a preference-aligned model can generate text that semantically aligns more closely with highly cited papers (referred to as *Chosen Texts*) than with less cited ones (*Rejected Texts*). Conceptually, this test evaluates the model's ability to internalize the linguistic and conceptual markers associated with future influence, such as clarity, novelty, or field relevance, and reproduce them in its own generative outputs. The experiment therefore serves as a proxy for assessing whether the learned representation captures not only discriminative preference boundaries but also constructive generative priors indicative of scholarly impact.

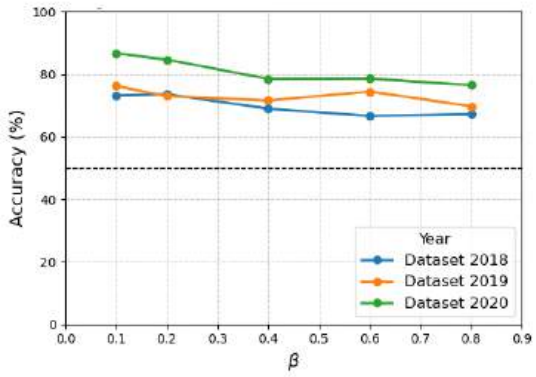

Fig. 7. **RQ4a:** Effect of β tuning on model accuracy across years for CS dataset. Lower values ($\beta = 0.1$) consistently yield superior performance, consistent with a low-gap preference regime. For readability, 95% bootstrap confidence intervals are omitted from this sensitivity plot; they are reported for the headline comparative accuracy figures and are available for the β analysis in the replication package.

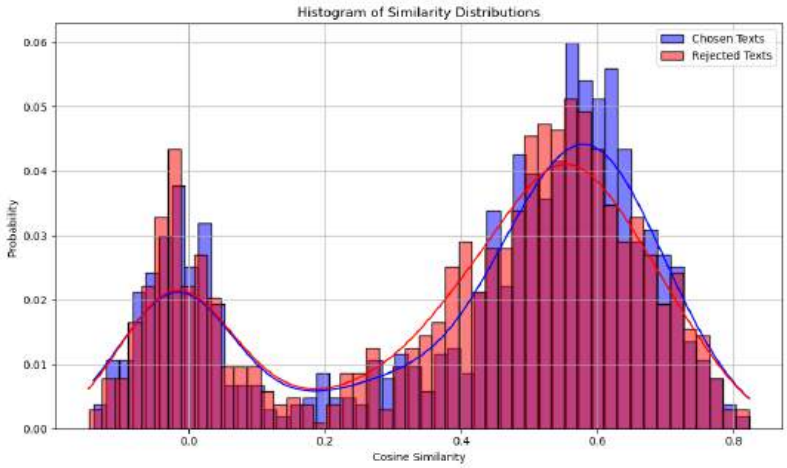

Fig. 8. **RQ4b:** Distributions of cosine similarity scores between generated texts and ground-truth abstracts for chosen (highly cited) and rejected (low-cited) papers.

To operationalize this evaluation, we prompted Impact-DPO on the 2020 test set to generate candidate abstracts conditioned on citing-paper context. Each generated sample was then compared to both the chosen and rejected ground-truth abstracts using cosine similarity computed over embeddings derived from the 'all-MiniLM-L6-v2' model [66]. This embedding-based comparison measures the degree to which generated text semantically approximates either outcome. A Kolmogorov–Smirnov (KS) test was conducted to verify whether the distributions of similarities between generated–chosen and generated–rejected pairs differ significantly. The KS statistic of 0.0744 ($p = 0.0065$) confirms that these distributions are statistically distinct, indicating that the model's outputs systematically align more closely with the language of highly cited papers.

As shown in Figure 8, the similarity distributions exhibit partial overlap but distinct modes. Generated texts show systematically higher similarity frequencies in the upper ranges of the chosen-paper distribution, whereas rejected-paper similarities are more diffuse and concentrated toward lower values. This pattern suggests that while the model can emulate general domain language, it preferentially reproduces linguistic and conceptual patterns characteristic of high-impact works. The residual overlap highlights the intrinsic difficulty of the task where rejected papers often share topical and stylistic features with their more cited counterparts, making absolute separation unlikely in a realistic academic corpus.

From a theoretical perspective, these findings indicate that preference alignment extends beyond classification accuracy to shape the model's generative manifold. By optimizing for pairwise citation preferences, the objective of Impact-DPO implicitly reweights latent linguistic dimensions associated with scholarly salience, such as the presence of specific methodological framing, field-specific terminology, or argument coherence. As a result, when asked to produce new text, the model tends to generate content that lies closer in semantic space to historically successful research outputs. This emergent generative alignment supports the broader hypothesis that preference optimization can encode normative patterns of influence within language generation.

7 Causal Interpretability of Learned Preferences

A central question for any preference-tuned model is *what textual features drive its comparative judgments*. High pairwise accuracy alone does not tell us whether Impact-DPO is responding to scientifically meaningful signals or merely exploiting superficial stylistic shortcuts. We therefore examine the learned preferences from several complementary perspectives: by intervening directly on textual concepts, by steering internal representations associated with those concepts, and by testing whether the model still discriminates meaningfully when citation gaps are small. The goal is not only to show that the model works, but also to articulate *why* it tends to prefer one abstract over another and where the evidence for that interpretation is strongest. All analyses are performed across both domains (CS and Physics) and all three cohort years (2018–2020), yielding six independent replication settings. The DPO model analyzed is LLaMA-2-7B fine-tuned with LoRA adapters (r=8, α=16, β=0.1).

We investigate four theoretically motivated textual features drawn from the science-of-science and scientific-communication literature [17, 59]: **(i)** *Conceptual Novelty*, operationalized as one minus the mean cosine similarity of a paper's SPECTER2 embedding to all same-domain embeddings published in or before that year [52, 60]; **(ii)** *Methodological Specificity*, measured as the density of pre-defined domain-specific technical keywords (matches per 100 words); **(iii)** *Clout* (rhetorical confidence), computed as the mean normalized Empath score over a LIWC-inspired category set (`dominant_personality`, `power`, `achievement`, `leader`, `competing`, `pride`, `strength`) [15, 40]; and **(iv)** *Analytic thinking*, computed as the mean normalized Empath score over `science`, `technology`, `intellectual`, `reading`, `writing`, `school`, and `math` [15, 40]. These features were selected because they correspond to plausible ingredients of perceived scholarly impact, i.e., whether a paper sounds new, whether it communicates methodological substance, and whether it presents its contribution with confident and analytical framing [1, 59, 60]. Abstract length is treated as a confounder throughout, since longer abstracts may appear more informative even when they are not genuinely more impactful [29, 59].

7.1 Counterfactual Text Interventions (RATE)

The most direct way to test whether a concept matters is to intervene on the text itself while keeping the scoring model fixed. We therefore implement the Rewrite-based Attribute Treatment Estimators (RATE) framework [16, 62]. For each test pair, we apply a double-rewrite procedure: (1) score the

original text x; (2) rewrite x to *remove* the target concept, yielding x'; and (3) rewrite x' again to *restore* the same concept, yielding x''. The resulting Causal Attribution Score is $\text{CAS} = r(x'') - r(x')$. Intuitively, this quantity asks how much the model's reward changes when the concept is toggled off and then back on, rather than merely correlating concept presence with higher scores.

This design is especially important because naive rewriting can introduce stylistic artifacts of its own. The double-rewrite construction reduces that problem by comparing two texts that have both passed through the same generative pipeline, differing primarily in the presence versus absence of the target concept. We further use Qwen 2.5-7B-Instruct as the rewriter while keeping LLaMA-2 as the DPO scorer, so that the intervention model is not simply learning to imitate the scorer's preferences. This cross-family setup makes it less likely that the measured effect is an artifact of a single model family optimizing for itself.

All 18 concept×domain-year conditions yield positive mean CAS, with 17/18 reaching statistical significance ($p < 0.05$); the sole exception is Physics 2020×Novelty ($p = 0.062$), whose confidence interval only narrowly crosses zero. Effects are larger in CS than in Physics (mean CAS $\approx$ 0.006–0.013 vs. 0.002–0.008), suggesting that the targeted concepts explain more of the decision boundary in CS. Importantly, CAS is predominantly negatively correlated with the original reward margin (ρ ranges from -0.32 to $+0.04$, with 16 of 18 conditions showing negative correlation). In practical terms, this means that novelty, specificity, and clout matter most when the model is deciding between difficult near-ties; when the preference signal is weak, adding or removing these concepts can tip the model from one side of the comparison to the other. This is stronger evidence than a simple feature-outcome association, because it shows that the model's score changes in the expected direction under controlled textual interventions.

7.2 Activation Steering via Representation Engineering

Counterfactual rewriting tells us that certain concepts causally affect the model's reward. A complementary question is whether those concepts are also reflected in the model's internal state. To probe this, we use activation steering via representation engineering [77]. For each concept, we construct 100 contrastive text pairs, extract hidden states from layers 15–20, and compute the first principal component of the positive–negative hidden-state difference to obtain a concept direction vector $\mathbf{v}_c \in \mathbb{R}^{4096}$. During inference on held-out test pairs (up to n=500, fewer where the test set is smaller), we add a scaled perturbation $\mathbf{h}'_l = \mathbf{h}_l + \alpha \cdot \mathbf{v}_c$ with $\alpha \in [-3, +3]$ and measure the resulting change in pairwise accuracy.

A concept direction should be interpreted as a linear axis in representation space along which the model becomes more or less sensitive to a given property, such as novelty or rhetorical confidence. Unlike RATE, steering does not alter the surface text; instead, it perturbs the hidden state and asks whether decision behavior changes in a concept-consistent way. This makes the test weaker as causal evidence than direct intervention on the text, but valuable as a mechanistic probe of whether the model internally organizes information along those dimensions.

Concept vectors are consistently found in layers 18–20, with PCA explained variance of 22–27% (Figure 9). Accuracy shifts of 0.6–1.8% across the steering range are modest but consistent across all 18 conditions without exception. The modest magnitude is expected where no single concept should determine the decision on its own, and the strongest effect should appear only for marginal pairs near the boundary. Taken together, the RATE and steering analyses paint a coherent picture. RATE provides the strongest support that novelty, specificity, and clout are causally relevant to the model's rewards, while activation steering suggests that information about those concepts is at least partially encoded in identifiable directions of the hidden state. The steering results should therefore be read as suggestive mechanistic evidence rather than as definitive proof of a single disentangled circuit.

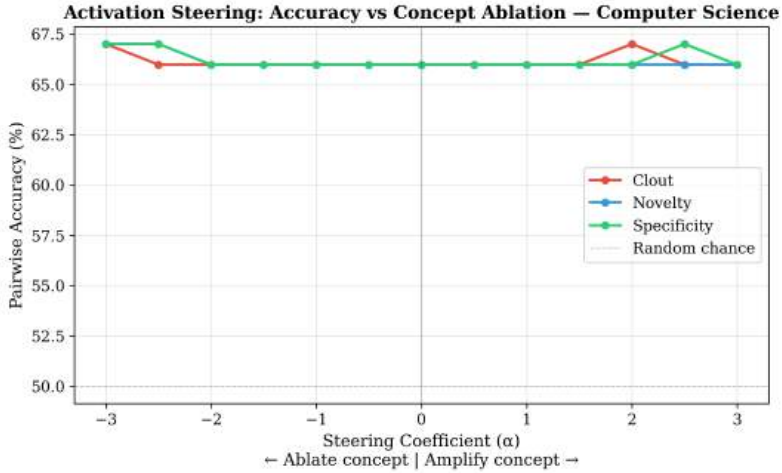

Fig. 9. Phase 3: Activation steering curves for CS 2020. Accuracy varies with steering magnitude α, suggesting that the model encodes novelty, specificity, and clout as partially separable directions in layers 19–20.

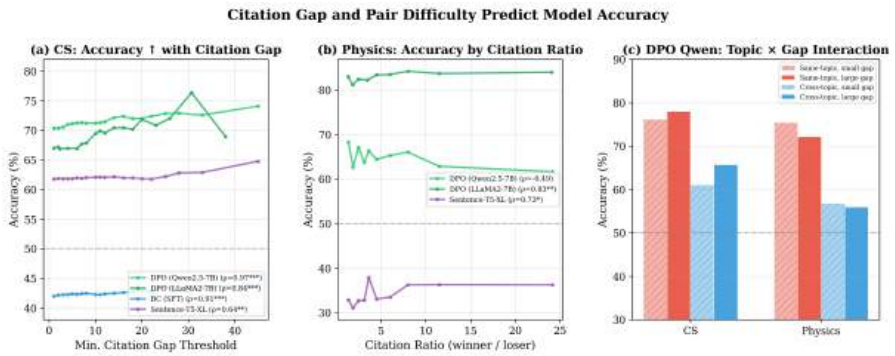

Fig. 10. Accuracy as a function of citation gap magnitude (decile-binned, all years pooled). Impact-DPO maintains above-chance accuracy even for low-gap pairs, and accuracy generally increases with gap size in the CS domain, confirming that the model captures genuine influence signals rather than relying solely on extreme outliers.

7.3 Citation Gap and the Scale of Impact

A further interpretive question concerns *where* the model succeeds. If high accuracy were driven only by very easy cases with extreme citation differences, then the learned preferences would be less informative about subtle scholarly judgment. We therefore ask whether the model's accuracy depends on the *magnitude* of the citation difference between paired papers, or whether it can still distinguish papers whose future influence diverges only slightly. To investigate this, we bin all test pairs by their absolute citation gap (deciles) and by their citation ratio, and compute Spearman rank correlations between gap magnitude and model accuracy.

Figure 10 reveals two key patterns. First, Impact-DPO achieves above-chance accuracy even for the lowest-gap quartile (CS pairs where the citation difference is ≤ 6, accuracy $\approx 67\%$), indicating that the model captures subtle influence signals beyond gross citation disparity. This is important

because these low-gap comparisons are precisely the cases in which the task most resembles a genuine preference judgment rather than a trivial separation between obvious winners and losers. Second, in the CS domain, accuracy generally increases with gap magnitude. When we compute accuracy for all pairs whose gap exceeds successive thresholds, Spearman $\rho = 0.84$ ($p < 0.001$) for the LLaMA-2 DPO model, confirming that the model's comparative judgment scales with the objective strength of the preference signal. In Physics, the relationship is weaker, likely because the heavier-tailed citation distribution (cf. Figure 1) compresses relative differences across deciles. Taken together, these findings address the concern that pairwise performance might be inflated by a small number of easy outliers. Instead, the model retains meaningful discriminative power on hard, low-gap pairs while still improving as the external preference signal becomes stronger.

7.4 Limitations of the Interpretability Analysis

Three limitations should be noted. First, all interpretability analyses were conducted on a single backbone (LLaMA-2-7B); whether the same features drive the preferences of the higher-performing Qwen 2.5-7B model remains an open question, and replicating the analysis on additional backbones is a priority for future work. Second, although RATE and activation steering provide stronger evidence than aggregate performance metrics alone, they are still partial probes of mechanism rather than exhaustive explanations. Appendix 8 therefore complements the aggregate analysis with four compact case studies spanning different difficulty regimes, but a broader set of case-level analyses would be needed to characterize failure modes systematically. In particular, the confident error case suggests that residual sensitivity to verbosity and assertive framing may still remain even when the higher-citation paper is substantively stronger. Third, our argument that co-citation pairing partially controls for cumulative-advantage confounds is theoretically motivated (shared citing context exposes both candidates to similar visibility conditions) but has not been verified experimentally—e.g., by measuring whether residual cumulative-advantage effects remain after conditioning on the co-citation structure. We acknowledge this as a theoretical rather than empirical claim.

8 Concluding Remarks

This paper introduced a preference-aligned framework for predicting and generating scholarly influence from textual and temporal cues in dynamic citation networks. Impact-DPO combines temporally informed prompting with direct preference optimization, enabling large language models to reason over evolving textual evidence without explicit graph message passing. Through systematic experiments across domains, years, and architectures, we demonstrated that this formulation provides a principled mechanism for aligning model predictions with the relative structure of scientific impact. The results show that preference-aligned models not only outperform embedding and prompting baselines on pairwise prediction tasks but also exhibit generative alignment, producing outputs that semantically resemble highly cited papers.

Our future research will extend this framework along three directions. (1) *Adaptive preference calibration.* The sensitivity of performance to the β parameter observed in RQ4a suggests the need for a dynamic calibration mechanism that adjusts alignment strength according to pairwise difficulty or temporal uncertainty. (2) *Cross-domain and structural generalization.* The stability of Impact-DPO across Computer Science and Physics (RQ2) motivates extending the approach to heterogeneous scholarly domains and incorporating explicit graph-based and author-level contextual signals to test the limits of cross-domain transfer. In particular, incorporating the citing author's own publication history and topical proximity to the cited work—capturing the citer's perspective—could further disentangle quality-driven from socially driven citation decisions. (3) *Generative forecasting of influence trajectories.* The emergent ability of the model to generate text semantically closer to

high-impact papers (RQ4b) opens the path toward controlled generation of prospective abstracts or hypotheses, enabling predictive simulation of future influential research.

References

[1] Honglin Bao and Misha Teplitskiy. 2024. A simulation-based analysis of the impact of rhetorical citations in science. *Nature Communications* 15, 1 (2024), 431.

[2] Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum. 2005. Time-aware authority ranking. *Internet Mathematics* 2, 3 (2005), 301–332.

[3] Clemens Blümel and Alexander Schniedermann. 2020. Studying review articles in scientometrics and beyond: a research agenda. *Scientometrics* 124, 1 (2020), 711–728.

[4] Michael D Brandt, Sherief A Ghozy, David F Kallmes, Robert J McDonald, and Ramanathan D Kadirvel. 2022. Comparison of citation rates between Covid-19 and non-Covid-19 articles across 24 major scientific journals. *PLoS One* 17, 7 (2022), e0271071.

[5] Peter Buneman, Dennis Dosso, Matteo Lissandrini, and Gianmaria Silvello. 2021. Data citation and the citation graph. *Quantitative Science Studies* 2, 4 (2021), 1399–1422.

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

[7] Zhangtao Cheng, Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Philip S Yu. 2024. Information cascade popularity prediction via probabilistic diffusion. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[8] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667* (2023).

[9] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

[10] Na Dai, Milad Shokouhi, and Brian D Davison. 2011. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 95–104.

[11] Joost CF de Winter. 2024. Can ChatGPT be used to predict citation counts, readership, and social media interaction? *Scientometrics* 129, 6 (2024), 3071–3087. https://doi.org/10.1007/s11192-024-05014-4

[12] Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics* 10 (2022), 257–273.

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *Advances in Neural Information Processing Systems*, Vol. 34.

[14] Mingli Ding, Wangke Yu, Tingyu Zeng, and Shuhua Wang. 2024. PTNS: patent citation trajectory prediction based on temporal network snapshots. *Scientific Reports* 14 (2024), 24034.

[15] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 4647–4657. https://doi.org/10.1145/2858036.2858535

[16] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, et al. 2022. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics* 10 (2022), 1–48.

[17] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, and Alessandro Vespignani. 2018. Science of science. *Science* 359, 6379 (2018), eaao0185. https://doi.org/10.1126/science.aao0185

[18] Giuseppe Giordano, Michelangelo Misuraca, and Marialuisa Restaino. 2025. Evaluating the speed of citation in scientific journals: A survival analysis-based approach. *Journal of Informetrics* 19, 3 (2025), 101714.

[19] Scott A. Hill. 2021. A measure for characterizing heavy-tailed networks. *Physical Review Research* 3, 2 (2021), 023257. https://doi.org/10.1103/PhysRevResearch.3.023257

[20] Minghao Hu, Yuxing Peng, and Xipeng Qiu. 2018. Attention-Guided Answer Distillation for Machine Reading Comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2077–2086.

[21] Ya-Han Hu, Chun-Tien Tai, Kang Ernest Liu, and Cheng-Fang Cai. 2020. Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics* 14, 1

(2020), 101004.
[22] Nourhan Ibrahim, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. 2024. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence* 4, 1 (2024), 76.
[23] Seyed Mehran Kazemi et al. 2020. Representation learning for dynamic graphs: A survey. *Journal of Machine Learning Research* 21, 70 (2020), 1–73.
[24] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. 2015. Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* 112, 24 (2015), 7426–7431.
[25] Qing Ke, Alexander J Gates, and Albert-László Barabási. 2023. A network-based normalized impact measure reveals successful periods of scientific discovery across disciplines. *Proceedings of the National Academy of Sciences* 120, 48 (2023), e2309378120.
[26] Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1317–1327.
[27] Habib Taha Kose, Jose Nunez-Yanez, Robert Piechocki, and James Pope. 2024. A survey of computationally efficient graph neural networks for reconfigurable systems. *Information* 15, 7 (2024), 377.
[28] Moreno La Quatra, Luca Cagliero, and Elena Baralis. 2021. Leveraging full-text article exploration for citation analysis. *Scientometrics* 126, 10 (2021), 8275–8293.
[29] Adrian Letchford, Tobias Preis, and Helen Susannah Moat. 2016. The Advantage of Simple Paper Abstracts. *Journal of Informetrics* 10, 1 (2016), 1–8. https://doi.org/10.1016/j.joi.2015.11.001
[30] Xin Li, Xiaodi Ma, and Ye Feng. 2024. Early identification of breakthrough research from sleeping beauties using machine learning. *Journal of Informetrics* 18, 2 (2024), 101517.
[31] Yiling Lin, Linzhuo Li, and Lingfei Wu. 2025. The disruption index measures displacement between a paper and its most cited reference. *arXiv preprint arXiv:2504.04677* (2025).
[32] Weidong Liu, Yu Zhang, Xiangfeng Luo, Yan Cao, Keqin Gan, Fuming Ye, Wei Tang, and Minglong Zhang. 2024. Patent transformation prediction: When a patent can be transformed. *Information Processing & Management* 61, 6 (2024), 103872.
[33] Yuang Liu, Wei Zhang, and Jun Wang. 2021. Adaptive Multi-Teacher Multi-level Knowledge Distillation. *Neurocomputing* 415 (2021), 106–113.
[34] Luca Mariotti, Veronica Guidetti, Federica Mandreoli, Andrea Belli, and Paolo Lombardi. 2024. Combining large language models with enterprise knowledge graphs: a perspective on enhanced natural language understanding. *Frontiers in artificial intelligence* 7 (2024), 1460065.
[35] Irina V. Marshakova. 1973. System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2* 6 (1973), 3–8.
[36] Mark EJ Newman. 2022. Ranking with multiple types of pairwise comparisons. *Proceedings of the Royal Society A* 478, 2266 (2022), 20220517.
[37] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. *arXiv preprint arXiv:2108.08877* (2021). https://arxiv.org/abs/2108.08877
[38] Michael Park, Erin Leahey, and Russell J Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature* 613, 7942 (2023), 138–144.
[39] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A Huberman, Kimmo Kaski, and Santo Fortunato. 2015. Attention decay in science. *Journal of Informetrics* 9, 4 (2015), 734–745.
[40] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015.* Technical Report. University of Texas at Austin, Austin, TX. https://doi.org/10.15781/T29G6Z
[41] Alexander M Petersen, Raj K Pan, Fabio Pammolli, and Santo Fortunato. 2019. Methods to account for citation inflation in research evaluation. *Research Policy* 48, 7 (2019), 1855–1865.
[42] Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's High Time: A Survey of Temporal Question Answering. *arXiv preprint arXiv:2505.20243* (2025).
[43] Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024. TimeR4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 6942–6952.
[44] Filippo Radicchi and Santo Fortunato. 2014. Power laws in citation distributions: evidence from Scopus. *Scientometrics* 101, 3 (2014), 1553–1564.
[45] Filippo Radicchi, Alexander Weissman, and Johan Bollen. 2017. Quantifying perceived impact of scientific publications. *Journal of Informetrics* 11, 3 (2017), 704–712.
[46] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, and Christopher D. Manning. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*.

[47] Andrew B Rosenkrantz, Ankur M Doshi, Luke A Ginocchio, and Yindalon Aphinyanaphongs. 2016. Use of a machine-learning method for predicting highly cited articles within general radiology journals. *Academic Radiology* 23, 12 (2016), 1573–1581.

[48] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Berthelot, Soroosh Mariooryad, Ye Tian, and Dung Bui. 2020. Adapting Models to Signal Degradation using Distillation. In *British Machine Vision Conference (BMVC)*.

[49] Tarek Saier, Johan Krause, and Michael Färber. 2023. unarxive 2022: All arxiv publications pre-processed for nlp, including structured full-text and citation network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 66–70.

[50] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.

[51] Gregor Schweiger, Adrian Barnett, Peter van den Besselaar, Lutz Bornmann, Andreas De Block, John P A Ioannidis, Ulf Sandström, and Stijn Conix. 2024. The costs of competition in distributing scarce research funds. *Proceedings of the National Academy of Sciences* 121, 52 (2024), e2407644121.

[52] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 5548–5566. https://doi.org/10.18653/v1/2023.emnlp-main.338 arXiv:2211.13228.

[53] Dylan Slack, Anna Hilgard, Sameer Singh, and Himanshu Lakkaraju. 2021. Considerations When Learning Additive Explanations for Black-Box Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[54] Henry Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24, 4 (1973), 265–269.

[55] Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072* (2024).

[56] Cecilia Summers and Michael J. Dinneen. 2021. On the Reproducibility of Neural Network Predictions. In *arXiv preprint arXiv:2102.03349*.

[57] Xigang Sun, Jingya Zhou, Ling Liu, and Wenqi Wei. 2023. Explicit time embedding based cascade attention network for information popularity prediction. *Information Processing & Management* 60, 3 (2023), 103278.

[58] Zhuanlan Sun, Weiye Chen, Xiangyu Li, Yanning Sun, Ruibo Wang, Shuhui Wang, and Yixing Fang. 2024. Textual features of peer review predict top-cited papers: An interpretable machine learning perspective. *Journal of Informetrics* 18, 2 (2024), 101501.

[59] Iman Tahamtan and Lutz Bornmann. 2019. What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics* 121, 3 (2019), 1189–1228. https://doi.org/10.1007/s11192-019-03243-4

[60] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science* 342, 6157 (2013), 468–472.

[61] Davide Vega and Matteo Magnani. 2018. Foundations of temporal text networks. *Applied network science* 3, 1 (2018), 25.

[62] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual Invariance to Spurious Correlations in Text Classification. In *Advances in Neural Information Processing Systems*, Vol. 34.

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

[64] Mingyang Wang, Guang Yu, Jianzhong Xu, Huixin He, Daren Yu, and Shuang An. 2012. Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics* 6, 4 (2012), 586–599.

[65] Tianduo Wang, Shichen Li, and Wei Lu. 2024. Self-training with direct preference optimization improves chain-of-thought reasoning. *arXiv preprint arXiv:2407.18248* (2024).

[66] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. *arXiv preprint arXiv:2012.15828* (2020). https://arxiv.org/abs/2012.15828

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

[68] Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large Language Models Can Learn Temporal Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 10452–10470. https://doi.org/10.18653/v1/2024.acl-long.563

[69] Rui Yan, Jie Tang, Xiaohua Liu, Dongdong Shan, and Xuanjing Li. 2011. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 1247–1252.

[70] Jinqing Yang, Wei Lu, Jiming Hu, and Shengzhi Huang. 2022. A novel emerging topic detection method: A knowledge ecology perspective. *Information Processing & Management* 59, 2 (2022), 102843.
[71] Fang Zhang and Shengli Wu. 2024. Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics* (2024). https://doi.org/10.1007/s11192-024-05086-0
[72] Zitong Zhang, Braja Gopal Patra, Ashraf Yaseen, Jie Zhu, Rachit Sabharwal, Kirk Roberts, Tru Cao, and Hulin Wu. 2023. Scholarly recommendation systems: a literature survey. *Knowledge and Information Systems* 65, 11 (2023), 4433–4478.
[73] Zhengang Zhang, Chuanming Yu, Jingnan Wang, and Lu An. 2025. A temporal evolution and fine-grained information aggregation model for citation count prediction. *Scientometrics* (2025), 1–23.
[74] Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. 2025. From Words to Worth: Newborn Article Impact Prediction with LLM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1183–1191.
[75] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Zhibo Chen, Ming Wen, and Wangmeng Zuo. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18697–18709.
[76] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).
[77] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* (2023).

Appendix A. Implementation, Environment, and Reproducibility

A.1 Hardware and System Environment

All experiments were executed on a single workstation configured as follows:

- **GPU:** NVIDIA RTX 6000 Ada (48 GB VRAM).
- **CPU:** 12 cores; **System RAM:** 250 GB.
- **CUDA/CuDNN stack:** CUDA 12.1/12.4 toolchain available; `pytorch-cuda =12.1`; cuDNN 9.1.0.70.
- **OS:** 64-bit Linux.

A.2 Core Libraries and Versions

We list the principal libraries that affect numerical behavior, optimization, and model loading; the full environment (hundreds of packages) is available upon request.

- **Python** 3.11.8; **PyTorch** 2.5.1; **TorchVision** 0.20.1; **Torchaudio** 2.3.0.
- **Transformers** 4.51.3; **TRL** 0.17.0; **PEFT** 0.17.2.dev0; **Accelerate** 0.30.1.
- **bitsandbytes** 0.42.0; **datasets** 3.0.0; **evaluate** 0.4.3.
- **scikit-learn** 1.5.1; **adapters** 1.2.0 (for SPECTER2).
- **sentence-transformers** 3.1.1; **Weights & Biases** 0.18.1.

A.3 Model-Specific Specifications

A.3.1 Preference-Aligned LLMs (DPO): Qwen 2.5–7B and LLaMA 2–7B.. We fine-tune causal LMs using Direct Preference Optimization (DPO) with LoRA adapters.

- **Backbones:** `Qwen/Qwen2.5-7B` and `meta-llama/Llama-2-7b-hf`.
- **LoRA:** target modules {`q_proj,k_proj,v_proj,o_proj`}; rank r=8; α=16; dropout 0.05; task type `CAUSAL_LM`. Gradient checkpointing enabled; `use_cache=False`.
- **Quantization & precision (from code defaults):** 4-bit `nf4` weight quantization via bitsandbytes with **bfloat16** compute (`bnb_4bit_compute_dtype=torch.bfloat16`). Attention implementation set to `eager`; FlashAttention 2 (`TRANSFORMERS_USE_FLASH_ATTENTION_2=false`).

- **Training schedule:** per-device batch size 4 (train) / 2 (eval); gradient accumulation 2; AdamW (paged, 32-bit) with learning rate 5×10^{-4}, cosine schedule, 100 warmup steps, weight decay 0.05.
- **DPO hyperparameters:** β=0.1 (globally used); evaluation and checkpointing every 2000 steps; best-checkpoint selection at end.
- **Sequence lengths:** max prompt length = 1024 tokens; max total length = 2048 tokens (truncation applied).
- **Data interface:** datasets loaded via `load_from_disk`; fields mapped to DPO triplets with `prompt` (temporal prompt), `chosen` (higher-future-citation paper), `rejected` (lower-future-citation paper).

A.3.2 SPECTER2 + SVR (Embedding Regressor).

- **Encoder:** `allenai/specter2_base` with regression adapter `allenai/specter2_regression` activated via the `adapters` library.
- **Regressor:** `LinearSVR` with grid search over $C \in \{10^{-4}, \ldots, 10^{2}\}$ (7 log-spaced values), 3-fold CV, `n_jobs=5`.
- **Pairwise decision:** scalar predictions for each candidate compared within pair.

A.3.3 Binary Classification (BC) Baseline.

- **Backbone:** runs used `sentence-transformers/sentence-t5-xl` (sequence classification head with `num_labels=2`); LoRA (`task_type=SEQ_CLS`) applied.
- **Precision: bfloat16** compute on GPU; no quantization.
- **Inputs:** two symmetric reformulations per pair (Choice 1/Choice 2 swapped) to reduce position bias; max sequence length 512.
- **Optimization:** AdamW, learning rate 5×10^{-4}, linear warmup (100 steps), gradient accumulation = 4.
- **Metrics:** accuracy and ROC–AUC; probabilities from softmax over logits.

Supplementary same-backbone BC check. To test whether the large DPO–BC gap could be explained by backbone capacity alone, we additionally fine-tuned `Qwen/Qwen2.5-7B` with LoRA for binary sequence classification with the same training hyperparameters as `Sentence-T5-XL`. Across the balanced held-out test sets, this same-backbone BC variant still remained near chance: CS 2018 55.64% [54.95, 56.34], CS 2019 58.08% [57.09, 59.02], CS 2020 51.90% [50.29, 53.55], Physics 2018 51.44% [50.69, 52.14], Physics 2019 53.65% [52.73, 54.57], and Physics 2020 50.02% [48.56, 51.50] (95% bootstrap CIs from 1 000 resamples). Physics 2020 is statistically indistinguishable from random guessing, and the full supplementary results are included in the replication package.

A.3.4 Prompt-Only Impact Scorers (NAIP-Style). We reproduce two prompt strategies commonly used in the literature.

- **Original (60-item) strategy.** The model outputs 60 integers in $[0, 100]$ over a fixed descriptor list; the mean is used as the score.
- **Improved (single-score) strategy.** The model outputs a single float in $[0, 1]$.
- **Providers.** Runs were conducted either via the official OpenAI SDK or a deterministic dummy provider for ablation; `NDCG@20` is computed when ground-truth TNCSI_SP is available.

A.4 Prompts for DPO and Pairwise Inference

DPO (training) prompt. As implemented in the training code:

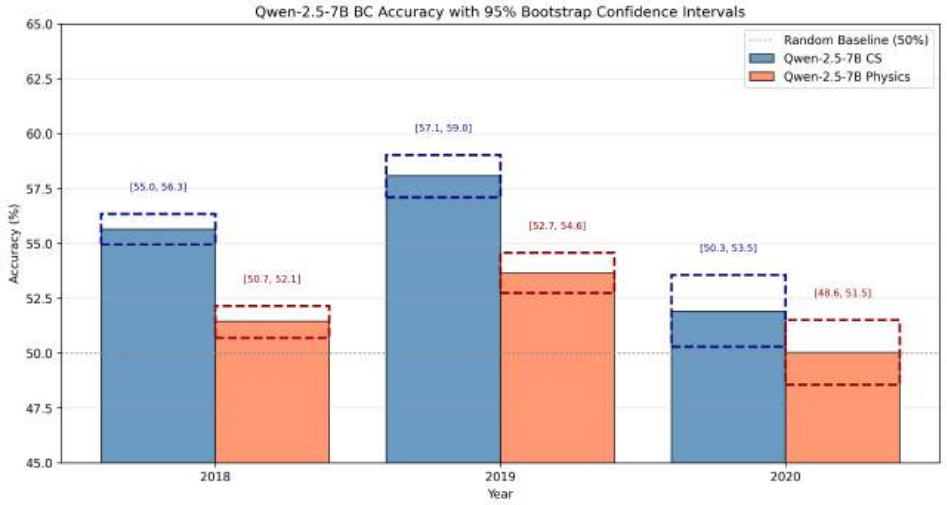

Fig. 11. **Supplementary same-backbone BC results with uncertainty.** Accuracy of Qwen 2.5–7B trained with a binary-classification objective across CS and Physics splits, with 95% bootstrap confidence intervals. All six runs remain near chance, and Physics 2020 is statistically indistinguishable from random guessing because its interval crosses the 50% baseline.

```
Question: Given the following abstract from year {q_year}, generate a highly
cited paper that cites this paper in year {rj_year}: {QUESTION_ABSTRACT}

Answer:
```

The `chosen` sequence corresponds to the higher-future-citation paper; `rejected` to the lower.

Clarification on training prompt direction. The DPO training prompt asks the model to "generate a highly cited paper that cites this paper." This phrasing may appear to reverse the prediction direction (generating a *citing* paper rather than choosing between *cited* papers). However, DPO does *not* generate text during training. Instead, it computes log-probability ratios of the `chosen` and `rejected` sequences under the current and reference policies. The prompt serves only as a contextual scaffold that situates the abstract in a citation-generation framing; the preference labels (chosen = higher-future-citation paper, rejected = lower) enforce the correct comparative direction. The generative framing was chosen deliberately: by asking the model to "generate" text conditioned on the citing abstract, the prompt elicits richer log-probability distributions over candidate abstracts than a purely discriminative phrasing (e.g., "which paper will be cited more?") would, because the model must assign token-level likelihoods to the full abstract rather than a single label token. At inference time, the pairwise prompt below is used, which explicitly asks the model to decide which candidate will accumulate more citations.

Pairwise (inference) prompt. For transparency and reproducibility, we provide a minimal deterministic template consistent with the paper's task definition; it returns a single-character decision:

> **Instruction.** You are given a citing paper and two papers it references. The citing paper was published in {YEAR_C}, Candidate A in {YEAR_A}, and Candidate B in {YEAR_B}. Decide which candidate will have more cumulative citations by the evaluation horizon T = {HORIZON}. Base your choice only on the scientific content and temporal cues below. **Answer with a single character: "A" or "B".**

Citing paper (year {YEAR_C}):
{CITING_TITLE}
{CITING_ABSTRACT}

Candidate A (year {YEAR_A}):
{A_TITLE}
{A_ABSTRACT}

Candidate B (year {YEAR_B}):
{B_TITLE}
{B_ABSTRACT}

Output: A **or** B.

NAIP-style prompts. We include the exact strings used:
Original (60-item) prompt (excerpt).

> Please rate the following abstract on each of the 60 items from 0 = Not at all to 100 = Very much. Only provide the numbers. For example:
> 1. 65 2. 50 3. 5 4. 95 5. . . .
> This is the abstract: {ABSTRACT}
> These are the items: 1. Engaging, 2. Controversial, . . . (full list of 60 descriptors).

Improved (single-score) prompt.

> Based on the following information, predict the academic impact of this paper as a single number between 0 and 1. Output only the number, with no additional text:
> Title: {TITLE}
> Abstract: {ABSTRACT}
> ONLY output a single number representing the future academic impact between 0 and 1. e.g., 0.69

A.5 Illustrative Case Studies

To provide case-level insight into *why* the model favours one abstract over another, we summarize four representative test pairs from the CS 2020 split in Table 3. The cases were selected to span different difficulty regimes: a high-confidence correct prediction, a low-margin correct prediction, a confident error, and a feature-driven correct prediction where length is nearly matched. The DPO implicit reward margin Δr is reported for each pair (positive = model prefers the higher-citation paper), together with lightweight feature proxies and a short qualitative interpretation.

Table 3 makes two patterns visible. First, the model's correct decisions often coincide with stronger novelty cues, clearer methodological substance, and more explicit contribution-oriented framing. Second, the confident error shows that length and rhetorical confidence can still mislead the model in some cases, which is why the interpretability claims in Section 7 should be read as informative but not exhaustive. Presenting the examples in tabular form keeps the comparison compact while still grounding the aggregate findings in concrete paper pairs.

A.6 Runtime and Throughput

On the RTX 6000 Ada (48 GB), DPO fine-tuning of Qwen 2.5–7B with 4-bit `nf4` weights and **bf16** compute (batch size 4, grad accumulation 2, max length 2048) runs comfortably without gradient OOM. The BC baseline (Sentence-T5-XL, bf16) is lightweight and does not require quantization. SPECTER2 embedding extraction is GPU-accelerated; SVR training is executed on CPU (12 cores) with 3-fold CV. Wall-clock hours per epoch depend on the split size and are stable across repeated runs with this configuration.

A.7 Reproducibility Notes

We control for common sources of nondeterminism:

- **Attention backend.** `attn_implementation=eager`; FlashAttention 2 disabled. This avoids backend-dependent variability.
- **Quantization.** All DPO runs use 4-bit `nf4` weights with **bf16** compute; BC uses fp16/bf16 without quantization; SPECTER2 uses fixed HF adapter weights.
- **Padding and truncation.** DPO: max prompt = 1024, max total length = 2048; BC: max length = 512.
- **Evaluation protocol.** All models are evaluated on identical, fixed pairwise test sets per domain-year. Scalar scorers (SPECTER2+SVR, NAIP-style) are converted to pairwise decisions via within-pair comparison.

Table 3. Four illustrative test pairs (CS 2020). "Chosen" denotes the higher-future-citation paper; "Rejected" the lower. Δr is the DPO implicit reward margin. Feature proxies: word count (Len), methodological-keyword density (Spec, per 100 words), novelty-keyword count (Nov), and assertive-language count (Clout). ✓/✗ indicates whether the model's pairwise prediction was correct.

Case	Higher-citation paper	Lower-citation paper	Result	Feature contrast	Interpretation
High-confidence correct	*Sequence-Level Knowledge Distillation* [26]	*Adapting Models to Signal Degradation using Distillation* [48]	$\Delta r = +0.111$ (✓)	Chosen: Len 181, Spec 3.9, Nov 5, Clout 1; Rejected: Len 122, Spec 2.5, Nov 2, Clout 2	The model strongly prefers the higher-impact paper when novelty and methodological detail are both clearly stronger and supported by concrete empirical framing.
Low-margin correct	*Adaptive Multi-Teacher Multi-level Knowledge Distillation* [33]	*On the Reproducibility of Neural Network Predictions* [56]	$\Delta r = +0.003$ (✓)	Chosen: Len 167, Spec 2.4, Nov 2, Clout 3; Rejected: Len 184, Spec 2.7, Nov 3, Clout 2	This near-tie suggests that, for marginal pairs, small differences in assertive framing and clarity of contribution can be enough to move the decision.
Confident incorrect	*CogView: Mastering Text-to-Image Generation via Transformers* [13]	*General Facial Representation Learning in a Visual-Linguistic Manner* (FaRL) [75]	$\Delta r = -0.058$ (✗)	Chosen: Len 84, Spec 3.6, Nov 2, Clout 1; Rejected: Len 153, Spec 0.7, Nov 3, Clout 2	The error illustrates a remaining failure mode: the model can overvalue longer, more rhetorically confident abstracts even when the truly higher-impact paper is shorter and more concise.
Feature-driven correct	*Attention-Guided Answer Distillation for Machine Reading Comprehension* [20]	*Considerations When Learning Additive Explanations for Black-Box Models* [53]	$\Delta r = +0.011$ (✓)	Chosen: Len 145, Spec 0.7, Nov 3, Clout 4; Rejected: Len 148, Spec 0.0, Nov 0, Clout 2	Because length is nearly matched, this case makes the preference signal cleaner: the model appears to separate the pair mainly through novelty and stronger contribution-oriented framing.