From Words to Worth: Newborn Article Impact Prediction with LLM

Penghai Zhao[1], Qinghua Xing[1], Kairan Dou[1], Jinyu Tian[1], Ying Tai[3], Jian Yang[1], Ming-Ming Cheng[12], Xiang Li[*12]

[1]VCIP, College of Computer Science, Nankai University
[2]NKIARI, Shenzhen Futian
[3]PCALab, Nanjing University
zhaopenghai@mail.nankai.edu.cn, xiang.li.implus@nankai.edu.cn

Abstract

Predicting the future impact of newly published articles is pivotal for advancing scientific discovery in an era of unprecedented scholarly expansion. This paper introduces a promising approach, leveraging the capabilities of LLMs to predict the future impact of newborn articles solely based on titles and abstracts. Breaking away from traditional methods heavily reliant on external data, we propose fine-tuning the LLM to uncover the intrinsic semantic patterns shared by highly impactful articles from a vast collection of text-score pairs. These semantic features are further utilized to predict the proposed indicator, $\mathrm{TNCSI}_{\mathrm{SP}}$, which incorporates favorable normalization properties across value, field, and time. To facilitate parameter-efficient fine-tuning of the LLM, we have also meticulously curated a dataset containing over 12,000 entries, each annotated with titles, abstracts, and their corresponding $\mathrm{TNCSI}_{\mathrm{SP}}$ values. Experimental results reveal an MAE of 0.216 and an NDCG@20 of 0.901, setting new benchmarks in predicting the impact of newborn articles. Finally, we present a real-world application example for predicting the impact of newborn journal articles to demonstrate its noteworthy practical value. Overall, our findings challenge existing paradigms and propose a shift towards a more content-focused prediction of academic impact, offering new insights for article impact prediction.

Link — sway.cloud.microsoft/KOH09sPR21Ubojbc

Demo — https://huggingface.co/spaces/ssocean/Newborn_Article_Impact_Predict

Introduction

The emerging field of article impact prediction is becoming increasingly critical in advancing scientific research. Generally, it focuses on forecasting the potential future citation counts of academic publications by exploiting the external data related to the article (Xia, Li, and Li 2023), such as early citation feature, venue characteristics, and author reputation, *etc*. Unlike traditional bibliometric evaluations that measure established influence, article impact prediction typically encompasses a broader range of applications. Large institutions utilize it for research funding decisions and academic promotions. Individuals may also benefit from impact predictions, which help them efficiently identify cutting-edge articles and remain leading in their fields, especially given the hundreds of daily arXiv submissions across various academic disciplines.

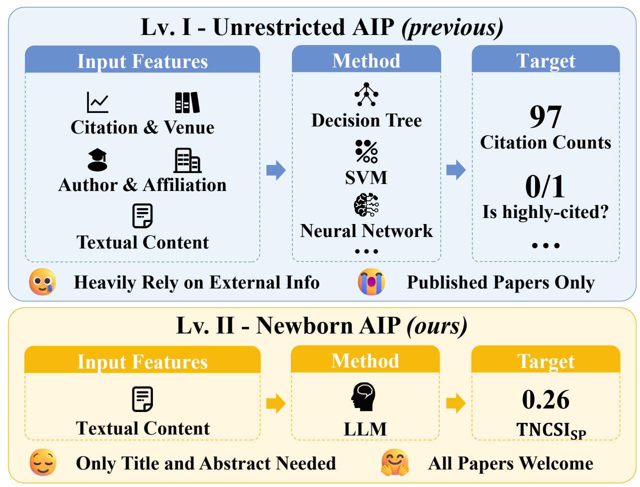

Figure 1: A taxonomy of article impact prediction (AIP): since there are virtually no other Lv. II methods, the "newborn AIP" segment represents the proposed approach, which predicts future academic impact in a "double-blind peer-review" manner.

Recently, as the field of Large Language Model (LLM) agent-based automated scientific research systems rapidly evolves (de la Torre-López, Ramírez, and Romero 2023; Wang et al. 2023; Lu et al. 2024), article impact prediction has never been more important than it is today. Imitating human experts, these autonomous systems typically start with identifying the most relevant and valuable research literature from extensive academic articles. Only then do these systems extract and synthesize knowledge from the retrieved literature, thereby enabling practical applications such as idea generation (Baek et al. 2024), and compound discovery (M. Bran et al. 2024), *etc*. As the saying goes, *You can't make bricks without straw*, article impact prediction has thus become a central component of automated research systems.

However, almost all existing impact prediction approaches rely on external historical data (Vergoulis et al. 2020; Wang et al. 2021; Zhao and Feng 2022; Abbas et al. 2023; Zhang and Wu 2024), which limits the practical value of these methods. Particularly, for those newly uploaded papers on pre-print websites (*e.g.*, arXiv), the absence of historical citation data and publication venue information poses challenges for existing methods to make accurate predictions. In addition, although most academic efforts prefer to predict citation counts, the validity of citation counts themselves remains debatable. As the Leiden Manifesto (Hicks et al. 2015) and the DORA Declaration (San Francisco 2018) indicate, citation count is not well suited for cross-disciplinary comparable evaluations and should not be used as the sole metric for assessing research impact. For example, a paper (Li et al. 2023) with one hundred citations might be unsurprising in the booming field of artificial intelligence, but in a relatively narrow yet equally important field such as paleontology, the paper with one hundred citations (Turner 2011) could be considered a cornerstone. Since automated scientific research systems mostly start from estimating the value of articles, such limitations undoubtedly weaken their ability to gather knowledge from other fields, thereby reducing the efficiency of knowledge synthesis.

To address the potential issues of regarding citation counts as the regression target, we first draw inspiration from the design principles of the Topic Normalized Citation Success Index (TNCSI) (Zhao et al. 2024) and make tailored improvements to adapt it for predicting the impact of newborn papers across various fields. The improved metric is named $\mathrm{TNCSI_{SP}}$, where SP stands for the Same Period, to highlight that the proposed metric is capable of comparing papers across varying time frames. Since key elements such as contribution, novelty, and insights are often reflected in the title and abstract, we contend that the "*worth*" of an article can typically be assessed by "*words*". Therefore, we try to regress the $\mathrm{TNCSI_{SP}}$ by feeding only the title and abstract to fine-tune the LLMs for reliable impact predictions.

For better clarity on the current state of impact prediction methods, we summarize and introduce a taxonomy based on the information required for the prediction (See in Fig. 1). The first level is called unrestricted article impact prediction, where predictions are permitted to rely on external historical information, and authors' reputation; this is the level at which most current methods are situated. The second level is named newborn article impact prediction, which particularly emphasizes making predictions about the impact only based on the article itself. This task is similar to a double-blind review process, where the model predicts the future impact without any author and affiliation information, publication details, or early citation data. Such an approach is particularly valuable for screening newly uploaded manuscripts, such as arXiv pre-prints and conference papers, as it may help researchers effectively identify the most promising articles. In this paper, we focus on the most challenging yet most valuable task: newborn article impact prediction.

To summarize, the core contributions of this work are as follows:

- **New Task**: We introduce a taxonomy and define a novel task entitled newborn article impact prediction, which aims to accurately predict the scholarly impact of newly published articles without external information.
- **New Method**: Tailored improvements have been made to the TNCSI, and for the first time, we demonstrate that fine-tuned LLMs are capable of predicting the future impact of newborn articles in a "double-blind review" setup.
- **New Dataset**: Accordingly, we have constructed and released the TKPD and NAID datasets. They are used to guide ChatGPT in generating topic key phrases and to train state-of-the-art LLMs for accurate article impact predictions, respectively.
- **Application**: Finally, we discuss and present an example of the proposed method's application in a real-world scenario, specifically in predicting the impact of journal articles published in 2024, with the hope of inspiring further advancements in the broader research field.

Related Work

Bibliometrics is a research field that utilizes quantitative analysis and statistical methods to assess the impact of scholarly publications. Typically, bibliometrics can be divided into two major categories: metrics for evaluating journals and metrics for evaluating individual articles. As the Leiden Manifesto (Hicks et al. 2015) and the DORA Declaration (San Francisco 2018) recommend, *do not use journal-based metrics to measure the quality of individual research articles*. Therefore, in this paper, we do not intend to use any journal-level bibliometric indicators (such as JIF (Garfield 1955)) as inputs or prediction targets. Instead, we focus on bibliometric indicators for individual articles. Tab. 1 illustrates the differences among them. Although FWCI and RCR are excellent metrics, their non-normalized numerical properties may impair the convergence of neural networks. Proposed by Zhao *et at.*, TNCSI (Zhao et al. 2024) features a clear physical meaning and favorable mathematical properties, representing the probability (ranges between 0 and 1) that an article's impact surpasses that of other articles in the same field. However, TNCSI is initially designed to evaluate review papers across different fields and is therefore not suitable for assessing normal research papers. Furthermore, TNCSI primarily focuses on assessing the existing impact of a review paper and does not normalize the impact of papers published in different years. This may lead to potential unfair comparisons in newborn article impact prediction tasks. Therefore, we propose an improved version to address the limitations of TNCSI. More details can be found in the Approach section.

Article Impact Prediction approaches typically adopt machine learning methods to forecast the future impact of articles. Most existing methods tend to exploit article statistical features, author characteristics, journal attributes, and historical citation data to aid decision trees, LSTM, MLP, and other machine learning algorithms in making predictions (Fu and Aliferis 2008; Wang, Yu, and Yu 2011; Qiu and Han 2024; Kousha and Thelwall 2024). Ruan *et al*. (Ruan et al. 2020) aims to enhance the prediction accuracy of five-year citation counts using a four-layer Back Propagation (BP) neural network by leveraging multiple features re-

Bibliometric	Value	Field	Time
Cites	×	×	×
FWCI (Colledge 2014)	×	✓	✓
RCR (Hutchins et al. 2016)	×	✓	✓
TNCSI (Zhao et al. 2024)	✓	✓	×
$\mathrm{TNCSI_{SP}}$ (*Ours*)	✓	✓	✓

Table 1: Several article-level bibliometrics for evaluating scholar Impact: value, field, and time respectively indicate whether the metric is a value between 0 and 1, whether it allows for cross-field comparisons, and whether it is suitable for the comparison of papers published at different times. These normalizations facilitate the network training.

lated to papers, journals, authors, references, and early citations. Ma *et al.* (Ma et al. 2021) propose a citation count prediction model that uses early citations and paper semantic features as input and employs Bi-LSTM for final predictions. Another notable citation-based machine learning approach exploits static features and time-dependent citation features to predict potentially excellent papers (Hu, Cui, and Lin 2023). In ABBAS's work (Abbas et al. 2023), an MLP-based method leveraging only external features is proposed to make prediction of future citation counts, achieving a decent performance with an NDCG of 0.95. Zhang et al. (Zhang and Wu 2024) discover that employing different models for papers in various domains significantly enhances the accuracy of prediction by leveraging early citation data. De (de Winter 2024) attempts to guide ChatGPT-4 in scoring over 2,000 paper abstracts from multiple perspectives, finding that the scores have Spearman correlation coefficients greater than 0.4 with Mendeley readership, and a correlation of 0.18 with citation counts. To the best of our knowledge, there is currently no method capable of accurately predicting the impact of an article based solely on the internal content.

Large Language Models have demonstrated powerful long-form text modeling capabilities and have been widely applied to various NLP tasks over the past few years, including dialogue systems, machine translation, sentiment analysis, *etc.* (Zhao et al. 2023; Tu et al. 2024; Jiang et al. 2024) Many commercial large language models (OpenAI 2022, 2023; Google 2024; Kimi.ai 2024) are not openly accessible, which prevents us from fine-tuning or instruction-tuning them. Therefore, we turned our attention to several excellent open-access large language models. LLaMA series (Touvron et al. 2023; AI 2024) are advanced language models created by Meta AI, available in multiple versions ranging from 7B to 70B parameters. It demonstrates decent performance on most tasks and has been widely adopted for various applications. Apart from LLaMA, there are several other notable open-source large language models, such as Qwen (Bai et al. 2023), Mistral (Jiang et al. 2023), Falcon (Almazrouei et al. 2023), *etc*. Regardless of the specific large language model, they were originally developed for autoregressive text generation. In this study, we use only the first generated token for numerical regression. Detailed descriptions and comprehensive evaluations of these models will be provided in the Approach and Experiment sections.

Approach

Tailored Improvement to the TNCSI

As mentioned in the Related Work section, TNCSI suffers from certain limitations, such as being restricted to evaluating review papers and only taking into account the cumulative impact of articles. We conducted a detailed analysis of its computational process and identified the reasons behind these limitations. First, TNCSI requires a predefined prompt template to guide ChatGPT in generating a corresponding review research area from the given title and abstract. The original prompt is specifically designed for review papers rather than normal research papers. Therefore, using their prompt directly on regular papers results in poor performance. Second, TNCSI primarily considers the cumulative impact of an article since its publication. However, for constructing datasets of the article impact prediction task, this approach may lead to potential issues of unfair comparison. Specifically, papers published earlier typically exhibit higher TNCSI values than recently published ones. This could potentially confuse the network's learning process, making LLM difficult to model the relationship between text features and its impact values.

Based on the discussion above, we make tailored improvements to the TNCSI and name the improved metric as $\mathrm{TNCSI_{SP}}$. Similar to the computational procedure of TNCSI, the procedure of the proposed $\mathrm{TNCSI_{SP}}$ is divided into three steps. In the first step, a well-designed prompt is utilized to guide ChatGPT (currently referred to gpt-3.5-turbo-0125) to identify the topic key phrase of an article. We have designed and tested a variety of prompt templates for identifying the article key phrase from different perspectives. To further mitigate individual cognitive biases, we enlisted the help of numerous researchers in the prompt creation process. All prompt templates are tested on a human-annotated dataset to evaluate the corresponding performance in the key phrase identification task. The second step involves using ChatGPT-generated key phrases to retrieve 1,000 related papers and their mata-info (*e.g.,* citation counts) from the Semantic Scholar API. Unlike TNCSI, which considers citation counts over the entire timeframe, $\mathrm{TNCSI_{SP}}$ focuses on the concurrent papers within a 6-month window before and after the publication date. This approach ensures that each paper is compared only to others published within a similar timeframe, thereby minimizing the citation advantage that older papers accumulate due to their extended presence. As a result, this method endows $\mathrm{TNCSI_{SP}}$ with the ability to normalize citation impact across different publication times. The final step remains consistent with TNCSI. The simplified mathematical expressions are shown as follows:

$$P(X = x) = \frac{\mathrm{Count}(p_x)}{C}. \tag{1}$$

Here, $C = 1000$ refers to the total number of retrieved papers. $\mathrm{Count}(p_x)$ represents the number of paper p with x citations. $P(X = x)$ is a discrete probability distribution

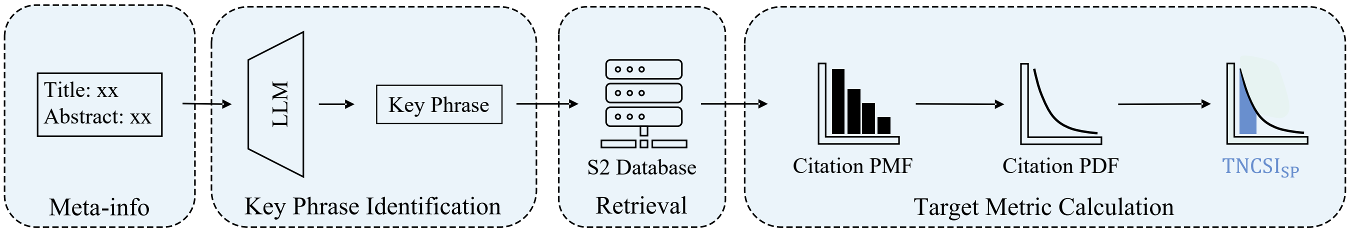

Figure 2: Flowchart for calculating $\mathrm{TNCSI_{SP}}$: $\mathrm{TNCSI_{SP}} \in [0, 1]$ represents the probability that a paper's citation count outperforms other papers in the same field and time period. "S2" refers to Semantic Scholar.

that describes the probability of a paper having exactly x citations among the retrieved C papers.

In their work (Zhao et al. 2024), $P(X = x)$ has been thoroughly discussed and is shown to follow an exponentially decaying distribution. Therefore, it could be converted into a probability density function using the maximum likelihood estimation. As shown in Eq. (2), we may derive the final $\mathrm{TNCSI_{SP}} \in [0, 1]$ by calculating the value of the corresponding definite integral:

$$TNCSI_{SP} = \int_0^{cites} \lambda e^{-\lambda x}\, dx, x \geq 0, \tag{2}$$

where $cites$ represents the number of citations that the paper being evaluated has received.

LLM for Newborn Article Impact Prediction

The autoregressive mechanism of large language models has been well-documented (Zhao et al. 2023). Essentially, these decoder-only models generate text in a sequential manner, with each token prediction relying on the context provided by the previous tokens. Such a paradigm allows it to fully leverage unlabeled data for self-supervised learning.

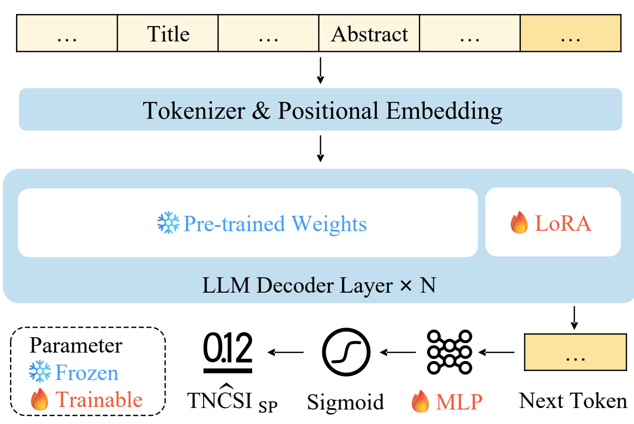

Figure 3: LLM as scholar impact predictor: overall framework of the proposed approach. Only the Next Token (first generated token) is used to regress the $\mathrm{TNCSI_{SP}}$.

In this paper, we maintain the autoregressive generation scheme of the large language model unchanged. However, unlike conventional text generation, we focus solely on the first token that the model generates autoregressively in response to user input. Specifically, assume the current input sequence is $\{w_1, w_2, \ldots, w_t\}$. The relationship between the LLM and the generation of the next token w_{t+1} can be expressed as:

$$w_{t+1} = \mathrm{LLM}(w_1, w_2, \ldots, w_t), \tag{3}$$

where $\mathrm{LLM}(\cdot)$ represents a large language model that predicts the next token in the sequence based on the input token sequence $\{w_1, w_2, \ldots, w_t\}$.

To facilitate the LLM's prediction of a single numerical value, we employ a simple multi-layer perceptron (MLP) to transform $w_{t+1} \in \mathbb{R}^{B\times 1\times D}$ into a real number $v \in \mathbb{R}$. Then, the value v is fed to a Sigmoid function, resulting in the predicted $\mathrm{T\hat{N}CSI_{SP}} \in [0, 1]$. Here, B represents the batch size, D is the dimension, and $\mathbb{R}$ denotes the set of real numbers. The process can be represented by the following equations:

$$T\hat{NCSI}_{SP} = \sigma(\mathrm{MLP}(w_{t+1})), \tag{4}$$

where $\mathrm{T\hat{N}CSI_{SP}}$ is computed by passing w_{t+1} through an MLP followed by a Sigmoid function σ.

Finally, we aim to minimize the mean square error (MSE) loss to align the predicted output $\mathrm{T\hat{N}CSI_{SP}}$ to the $\mathrm{TNCSI_{SP}}$ obtained from previous statistical calculations.

In practice, the immense number of parameters in large language models requires substantial computational resources for training, which exceeds our practical capacity. Therefore, we adopted low-rank matrix decomposition (LoRA) (Hu et al. 2021) and model quantization techniques (Dettmers et al. 2022) to reduce computational resource consumption and accelerate network training and inference processes. We recommend readers refer to the original papers for further details.

Datasets Construction

We have constructed a total of two datasets, the Topic Key Phrase Dataset (TKPD) and the Normalized Article Influence Dataset (NAID). Each of these datasets serves different purposes, which will be described in more detail below.

Topic Key Phrase Dataset: TKPD includes 251 entries encompassing titles, abstracts, and core task or field names of random articles across various fields in artificial intelligence. To mitigate subjectivity in our study and ensure consistent annotations, we employ manual labeling of key phrases by a seasoned AI researcher and enlist three additional researchers to double-check the annotations. Due to the specialized knowledge required for data annotation, this paper is precluded from annotating articles from non-AI fields. Nevertheless, we believe that the inconsistencies between the different subfields within the AI field are sufficient to simulate the differences between the distinct disciplines.

Normalized Article Impact Dataset: NAID is used to train LLMs to predict the impact of articles. It comprises the title, abstract, and the corresponding $\mathrm{TNCSI_{SP}}$, *etc*. The NAID consists of over 12,000 data entries from various AI fields, excluding survey papers, and includes papers with category IDs "cs.CV", "cs.CL", and "cs.AI" uploaded to arXiv between 2020 and 2022. In particular, the "cs.AI" category spans a broad range of disciplines such as mathematics, physics, and cognitive science, thereby extending the training data beyond the AI domain. NAID is a uniformly distributed dataset, meaning that the sources of the papers, the original publication year of the papers, and the corresponding $\mathrm{TNCSI_{SP}}$ values are evenly distributed.

Experiments

Metrics

Mean Absolute Error: MAE is employed to assess the prediction accuracy. It is a measure used to evaluate the discrepancy between the predicted value and the ground truth y_i, which is defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5}$$

where n represents the number of samples in the test set,y_i denotes the actual output (*e.g.*, $\mathrm{TNCSI_{SP}}$), $\hat{y}_i$ stands for the predicted value (*e.g.*, $\hat{TNCSI}_{SP}$), $|y_i - \hat{y}_i|$ is the absolute difference between the actual and predicted values. Generally, a lower MAE indicates a higher accuracy of the model's predictions.

Normalized Discounted Cumulative Gain (Järvelin and Kekäläinen 2000): NDCG is another metric to evaluate the effectiveness of the prediction. NDCG, originally developed for recommendation systems to measure the gain of a document based on its position in the recommended list, is calculated as follows:

$$\mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}, \tag{6}$$

where DCG@K = $\sum_{i=1}^{K}(2^{\hat{y}_i} - 1)/\log_2(i+1)$, and IDCG@K = $\sum_{i=1}^{K}(2^{y_i} - 1)/\log_2(i+1)$. $K = 20$ represents the position cutoff for the recommended list. NDCG is a metric that ranges from 0 to 1, with scores closer to 1 reflecting that more influential documents are ranked higher, signifying better performance.

Normalized Edit Distance (Yujian and Bo 2007): NED is a metric to measure the similarity between two strings by normalizing the edit distance by the length of the longer string. It is defined as:

$$\mathrm{NED}(A,B) = \frac{\mathrm{ED}(A,B)}{\max(|A|,|B|)}, \tag{7}$$

where $\mathrm{ED}(A,B)$ is the edit distance between strings A and B, and $\max(|A|,|B|)$ is the length of the longer string. The lower the NED value, the more similar the two strings are.

Comparison with Previous Methods

We employ NDCG to assess the effectiveness of different methods in identifying high-impact papers, given their varying prediction targets. To ensure fairness, we exclude the external data relied upon by previous approaches. Reproduction details are provided in the Appendix.

Tab. 2 clearly illustrates the performance differences between our proposed method and previous methods in identifying potentially high-impact papers. The proposed method demonstrates a notable superiority in the newborn article impact prediction setting compared to earlier representative works. Most level I methods underperform without external information. For example, the LSTM-based method (Ma et al. 2021) reports an NDCG of 0.84 when leveraging external information, but its performance drops significantly to 0.196 when relying solely on the title and abstract, suggesting a limited capacity to effectively map semantic features to the target $\mathrm{TNCSI_{SP}}$. The less favorable performance of ChatGPT-generated and LLaMA-3-generated approaches in identifying high-impact papers suggests that zero-shot LLM-generated approaches still require further exploration. In summary, we believe the significant performance improvements of the proposed approach could be credited to LLaMA-3's extensive foundational knowledge and the incorporation of the $\mathrm{TNCSI_{SP}}$ metric during fine-tuning, which enhances its ability to identify impactful semantic features across various domains and time periods.

Performance of Various LLMs

As the central task of this paper, we comprehensively evaluate the performance of various LLMs on the NAID test set. As shown in Tab. 3, LLaMA-3-8B achieves the best overall performance. Interestingly, we observe that MAE and NDCG are not always inversely correlated; for example, while Falcon achieves a lower MAE than Phi-3, its NDCG is slightly lower. This suggests that Falcon is more accurate in predicting lower-impact papers but less effective for high-impact ones. Since our primary focus is on identifying high-impact papers, a higher NDCG is generally more advantageous than a lower MAE in this scenario.

The Qwen family is selected to further explore the effects of model size on the performance of article impact prediction tasks. Compared to the LLaMA series, the Qwen series features more official models with smaller parameter sizes, specifically 0.5B, 1.5B, and 7B. We train each of these models on the NAID train set, and the test results are illustrated in Fig. 4. It can be observed that as the model parameter size increases, the performance correspondingly improves.

Effectiveness of Prompt Engineering

This paper conducts prompt engineering on two tasks: identifying the topic key phrase for calculating $\mathrm{TNCSI_{SP}}$, and guiding the LLM to make predictions.

For Identifying Key Phrase: We conduct numerous experiments to test the performance of different models and prompts on the TKPD. Tab. 4 reports the NED for 3 representative prompts on the TKPD. Ultimately, we employ the

Methods	Ori. Lv.	Input Feature for Fair Comparison	Target	NDCG ↑
MLP-based (Ruan et al. 2020)	I	paper length, reference numbers, *etc.*	Cites	0.147
LSTM-based (Ma et al. 2021)	I	title, and abstract	Cites	0.196
Model Ensemble (Zhang and Wu 2024)	I	(Ruan et al. 2020) + research filed	Cites	0.201
MLP-based (Hu, Cui, and Lin 2023)	I	the same to (Ruan et al. 2020)	Is Top 5%	0.464
ChatGPT-generated (de Winter 2024)	II	title, and abstract	Score	0.597
LLaMA-3-generated	II	title, and abstract	TNCSI_{SP}	0.674
Fine-tuned LLaMA-3-based	II	title, and abstract	Cites	0.403
Fine-tuned LLaMA-3-based	II	title, and abstract	FWCI	0.594
Fine-tuned LLaMA-3-based (*ours*)	II	title, and abstract	TNCSI_{SP}	**0.901**

Table 2: Comparison with previous approaches: the proposed method achieves notable advantages among others. An upward arrow indicates that a higher value is better, and vice versa. Bold font denotes the best performance among all methods. 'Ori. Lv.' refers to the taxonomy level of the original study, while 'Target' denotes the predicted target type in the corresponding research. See the Appendix for more reproduction details.

LLMs	Size ↓	MAE ↓	NDCG ↑	Memory ↓
Phi-3	**3.8B**	0.226	0.742	**6.2GB**
Falcon	7B	0.231	0.740	8.9GB
Qwen-2	7B	0.223	0.774	12.6GB
Mistral	7B	0.220	0.850	15.4GB
LLaMA-3	8B	**0.216**	**0.901**	9.4GB

Table 3: Performance comparison of different LLMs on the NAID test set: "Memory" stands for the minimum memory usage during inference.

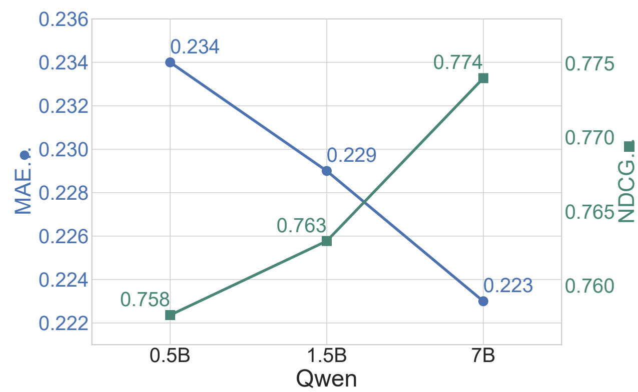

Figure 4: The impact of various model parameters on performance: the larger the number of model parameters, the better the performance.

prompt template from the last row, along with gpt-3.5-turbo-0125, to generate topic key phrases. Extended experimental records can be found in the Supplementary Material.

User Prompt Template	NED↓
Identify the research field from the given title and abstract. You MUST respond with the keyword ONLY in this format: xxx	0.30
Based on the title and abstract, determine the main area of study for the paper, focusing on a keyword that accurately represents the field. You MUST respond with the keyword ONLY in this format: xxx.	0.29
Given the title and abstract below, determine the specific research field by focusing on the main application area and the key technology. You MUST respond with the keyword ONLY in this format: xxx.	**0.26**

Table 4: Comparison of various user prompts for identifying topic key phrase.

For Guiding LLM: As shown in Tab. 5, We test several prompt templates to wrap the title and abstract before inputting them into the fine-tuned LLM. Despite PEFT, variations in prompt templates affect performance; more detailed descriptions often lead to better results. However, overly detailed prompts may also cause a slight NDCG decrease.

Comparative Analysis of the TNCSI_{SP}

We have trained LLaMA-3 on articles from different years and with various regression targets to demonstrate the superiority of the proposed TNCSI_{SP}. As shown in Fig. 5, when targeting the improved TNCSI_{SP}, the model provides more stable predictions regarding articles from different years. Table 6 further demonstrates the generalizability of TNCSI_{SP}. It enables various types of models to better resist the bias accumulated over time. This suggests that TNCSI_{SP} empowers the model to identify semantic features shared by high-impact articles across different years, thereby achieving significant improvements in overall task performance.

Applications

In this section, we present an intriguing example, journal average impact prediction, to further demonstrate the effectiveness of our method in real-world applications.

Theoretically, journals in different quartiles are expected to exhibit varying average impacts. Therefore, we guide the LLaMA-3 to predict the average TNCSI_{SP} of articles

Prompt Template	NDCG↑
Title: {title} \n Abstract: {abstract}.	0.849
Given the provided title and abstract, predict the future normalized academic impact on a scale from 0 (lowest impact) to 1 (highest impact). You may consider factors such as the language clarity, novelty of the research, or the claim of state-of-the-art, etc. Title: {title}\nAbstract: {abstract}	0.869
Given a certain paper entitled {title}, and its abstract: {abstract}. Predict its normalized scholar impact:	0.889
Given a certain paper entitled {title}, and its abstract: {abstract}. Predict its normalized scholar impact (between 0 and 1):	**0.901**

Table 5: Comparison of various prompts for guiding LLMs to predict the future impact.

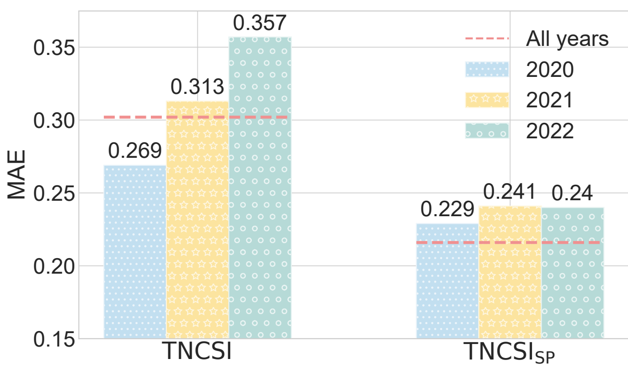

Figure 5: Impact of different prediction targets on performance: $\mathrm{TNCSI_{SP}}$ demonstrates superior performance over TNCSI with training data from different years.

published in 2024 across several journals from different JCR quartiles within the field of computer science. Since LLaMA-3's training data only extends up to early 2023, it is highly unlikely that the model has encountered these articles, significantly reducing the risk of data leakage. It is also worth noting that a journal's impact factor would be significantly influenced by a small number of highly cited papers (Lei and Sun 2020; Leydesdorff 2012). To this end, we analyzed over 500 randomly selected articles from various journals across different quartiles for impact prediction, focusing on the average predicted $\mathrm{TNCSI_{SP}}$ of the top 5% and 25% of notable papers within each quartile. In Fig. 6, we observe a clear positive correlation between the predicted impact of the notable top 5% of articles and their respective quartiles. Although the predicted impact of the top 25% of articles in the Q2 quartile is slightly higher than that of Q1, it is still considered a reasonable phenomenon.

Beyond journal impact prediction, our system holds promise for a variety of other real-world applications. For instance, given the vast number of daily pre-print submissions, the proposed approach may also help efficiently identify high-quality research worth closer examination. It may significantly reduce the time researchers spend reviewing large volumes of papers on arXiv, thereby enhancing overall research efficiency.

Methods	Target	NDCG ↑
MLP-based	TNCSI	0.464
MLP-based	TNCSI$_{SP}$	0.634
LSTM-based	TNCSI	0.373
LSTM-based	TNCSI$_{SP}$	0.646
Fine-tuned LLaMA-3-based	TNCSI	0.865
Fine-tuned LLaMA-3-based	TNCSI$_{SP}$	**0.901**

Table 6: Performance comparison using the TNCSI$_{SP}$ metric: all methods show improvements when targeting TNCSI$_{SP}$. The input is consistent with Tab. 2.

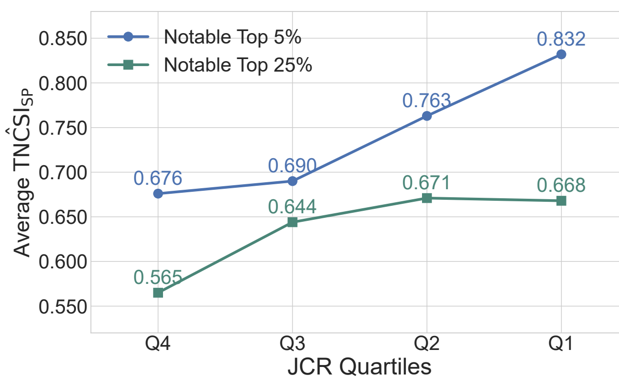

Figure 6: Predicted $\mathrm{TNCSI_{SP}}$ values for journals in different JCR quartiles: higher quartiles show higher predicted values. To avoid potential conflicts of interest, we denote Q1, Q2, Q3, and Q4 to represent articles from journals in JCR quartiles 1, 2, 3, and 4, respectively.

Conclusion

In this paper, we demonstrate the potential of LLMs for predicting the tailored $\mathrm{TNCSI_{SP}}$ of newborn papers with titles and abstracts only. The NAID dataset, comprising over 12,000 entries, is constructed and utilized to fine-tune various advanced LLMs. Empirical evaluations demonstrate that the LLaMA-3 model, with an MAE of 0.216 and an NDCG@20 of 0.901, significantly surpasses the performance of prior methods when solely relying on internal information. Furthermore, the impact values predicted by our method show a strong positive correlation with the quartile rankings of journals for articles published in 2024, illustrating the practical applicability of our approach in real-world settings. Overall, the proposed approach effectively estimates a future impact score from 0 to 1 for newly published papers, presenting considerable benefits for individuals, institutions, and automated scientific research systems.

Ethical Statement

We are aware of the potential for manipulation through excessive optimization of titles and abstracts. Researchers must refrain from excessively embellishing titles and abstracts, particularly by making false claims about unachieved performance or overly exaggerating the significance of their methods, in an attempt to manipulate predicted impact values.

Due to constraints such as the access frequency limits of the Semantic Scholar API, we are unable to construct a larger dataset. Therefore, our proposed method only serves as a preliminary exploratory approach. The predictions generated by this method are probabilistic estimates and should never be considered definitive assessments of an article's quality. The method is intended to provide additional insights and must not replace the existing peer-review process, which remains essential for maintaining the integrity and rigor of academic research. The authors are not responsible for any decisions made based on the predictions.

Acknowledgements

This research was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206134), the Fundamental Research Funds for the Central Universities 070-63233084, and the Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP). Computation is supported by the Supercomputing Center of Nankai University (NKSC). This work was supported by the National Science Fund of China under Grant No. 62361166670.

References

Abbas, K.; Hasan, M. K.; Abbasi, A.; Mokhtar, U. A.; Khan, A.; Abdullah, S. N. H. S.; Dong, S.; Islam, S.; Alboaneen, D.; and Ahmed, F. R. A. 2023. Predicting the future popularity of academic publications using deep learning by considering it as temporal citation networks. *IEEE Access*.

AI, M. 2024. Meta LLaMA 3: Advancements in Open-Source Large Language Models. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-04.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, É.; Hesslow, D.; Launay, J.; Malartic, Q.; et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Baek, J.; Jauhar, S. K.; Cucerzan, S.; and Hwang, S. J. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Colledge, L. 2014. Snowball metrics recipe book. *Amsterdam: Snowball Metrics Program Partners*, 110: 82.

de la Torre-López, J.; Ramírez, A.; and Romero, J. R. 2023. Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10): 2171–2194.

de Winter, J. 2024. Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, 1–19.

Dettmers, T.; Lewis, M.; Belkada, Y.; and Zettlemoyer, L. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35: 30318–30332.

Fu, L. D.; and Aliferis, C. 2008. Models for predicting and explaining citation count of biomedical articles. In *AMIA Annual symposium proceedings*, volume 2008, 222. American Medical Informatics Association.

Garfield, E. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159): 108–111.

Google. 2024. Gemini: Google's AI Model. https://www.google.com/gemini. Accessed: 2024-07-04.

Hicks, D.; Wouters, P.; Waltman, L.; De Rijcke, S.; and Rafols, I. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, 520(7548): 429–431.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Z.; Cui, J.; and Lin, A. 2023. Identifying potentially excellent publications using a citation-based machine learning approach. *Information Processing & Management*, 60(3): 103323.

Hutchins, B. I.; Yuan, X.; Anderson, J. M.; and Santangelo, G. M. 2016. Relative citation ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS biology*, 14(9): e1002541.

Järvelin, K.; and Kekäläinen, J. 2000. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, 243–250. ACM New York, NY, USA.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, Y.; Yan, X.; Ji, G.-P.; Fu, K.; Sun, M.; Xiong, H.; Fan, D.-P.; and Khan, F. S. 2024. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1): 17.

Kimi.ai. 2024. Kimi.ai. https://kimi.moonshot.cn/. Accessed: 2024-07-04.

Kousha, K.; and Thelwall, M. 2024. Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 75(3): 215–244.

Lei, L.; and Sun, Y. 2020. Should highly cited items be excluded in impact factor calculation? The effect of review articles on journal impact factor. *Scientometrics*, 122(3): 1697–1706.

Leydesdorff, L. 2012. Alternatives to the journal impact factor: I3 and the top-10%(or top-25%?) of the most-highly cited papers. *Scientometrics*, 92(2): 355–365.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16794–16805.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.

M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1–11.

Ma, A.; Liu, Y.; Xu, X.; and Dong, T. 2021. A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, 126(8): 6803–6823.

OpenAI. 2022. ChatGPT: optimizing language models for dialogue. https://openai.com/blog/chatgpt/. Accessed 25 December 2023.

OpenAI. 2023. Generative Pre-trained Transformer 4 (GPT-4). https://openai.com/gpt-4/. Accessed 25 December 2023.

Qiu, J.; and Han, X. 2024. An Early Evaluation of the Long-Term Influence of Academic Papers Based on Machine Learning Algorithms. *IEEE Access*, 12: 41773–41786.

Ruan, X.; Zhu, Y.; Li, J.; and Cheng, Y. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, 14(3): 101039.

San Francisco, D. 2018. San Francisco declaration on research assessment.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, X.; He, Z.; Huang, Y.; Zhang, Z.-H.; Yang, M.; and Zhao, J. 2024. An Overview of Large AI Models and Their Applications. *Visual Intelligence*, 2.

Turner, D. 2011. *Paleontology: a philosophical introduction*. Cambridge University Press.

Vergoulis, T.; Kanellos, I.; Giannopoulos, G.; and Dalamagas, T. 2020. Simplifying Impact Prediction for Scientific Articles. *ArXiv*, abs/2012.15192.

Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47–60.

Wang, K.; Shi, W.; Bai, J.; Zhao, X.; and Zhang, L. 2021. Prediction and application of article potential citations based on nonlinear citation-forecasting combined model. *Scientometrics*, 126: 6533–6550.

Wang, M.; Yu, G.; and Yu, D. 2011. Mining typical features for highly cited papers. *Scientometrics*, 87(3): 695–706.

Xia, W.; Li, T.; and Li, C. 2023. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, 128(1): 543–585.

Yujian, L.; and Bo, L. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6): 1091–1095.

Zhang, F.; and Wu, S. 2024. Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics*, 1–30.

Zhao, P.; Zhang, X.; Cheng, M.-M.; Yang, J.; and Li, X. 2024. A Literature Review of Literature Reviews in Pattern Analysis and Machine Intelligence. *arXiv preprint arXiv:2402.12928*.

Zhao, Q.; and Feng, X. 2022. Utilizing citation network structure to predict paper citation counts: A deep learning approach. *Journal of Informetrics*, 16(1): 101235.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.