从文字到价值：利用 LLM 预测新生文章影响力

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

在学术研究不断涌现的时代，预测新发表文章的未来影响力对于推动科学发现至关重要。本文提出了一种有前景的方法，引导大语言模型仅通过标题和摘要来预测新文章的未来影响力。不同于传统方法过度依赖外部数据，我们提出对大语言模型进行微调，以揭示从大量文本-评分对中共享的具有高度影响力文章的内在语义模式。这些语义特征进一步用于预测提出的指标 TNCSI_{SP}，该指标在值、领域和时间方面具备有利的归一化特性。为了便于参数高效微调，我们还精心策划了一个包含超过 12,000 条样本的数据集，每条条目都标注了标题、摘要及其对应的 TNCSI_{SP} 值。实验结果显示，MAE 为 0.216，NDCG@20 为 0.901，为预测新生文章影响力任务设定了新的基准。最后，我们展示了一个实际应用示例，用于预测新生期刊文章的影响力，以展示其显著的实际价值。总体而言，我们的研究结果挑战了现有范式，呼吁转向更加以内容为中心的学术影响力预测，提供了对文章影响力预测的新见解。

Link — sway.cloud.microsoft/KOH09sPR21Ubojbc

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

Introduction

文章影响力预测这一新兴领域在推动科学研究方面正变得越来越重要。一般来说，它主要是通过利用与文章相关的外部数据 (Xia, Li, and Li 2023)，如早期引用特征、发表地点特征和作者声誉等，预测学术出版物未来的潜在引用量。与衡量既定影响力的传统文献计量学评价不同，文章影响力预测的应用范围更广。大型机构将其用于研究经费决策和学术晋升。个人也可以从影响力预测中获益，因为它可以帮助他们有效地识别前沿文章，并保持在各自领域的领先地位，尤其是考虑到每天有数以百计的 arXiv 投稿横跨各个学术学科。

最近，随着基于大语言模型代理的自动化科研系统领域的快速发展 (de la Torre-López, Ramírez, and Romero 2023; Wang et al. 2023; Lu et al. 2024)，文章影响力预测从未像今天这样重要。这些自主系统模仿人类专家，通常首先从大量学术文章中识别出最相关、最有价值的研究文献。然后，这些系统才会从检索到的文献中提取和综合知识，从而实现实际应用，如创意生成（idea generation）(Baek et al. 2024) 和化合物发现（compound discovery）(M. Bran et al. 2024) 等。正所谓 “巧妇难为无米之炊”，文章影响力预测已经成为自动化研究系统的核心组成部分。

然而，几乎所有现有的影响力预测方法都依赖于外部历史数据 (Vergoulis et al. 2020; Wang et al. 2021; Zhao and Feng 2022; Abbas et al. 2023; Zhang and Wu 2024)，这限制了这些方法的实用价值。特别是对于那些新上传至预发表网站（如 arXiv）的论文，由于缺乏历史引用数据和发表地点信息，现有方法很难做出准确预测。此外，尽管大多数学术界都倾向于预测引用次数，但引用次数本身的有效性仍有待商榷。正如《莱顿宣言》（Leiden Manifesto）(Hicks et al. 2015) 和《DORA 宣言》（DORA Declaration）(San Francisco 2018) 所指出的，引用次数并不适合用于跨学科可比性对比，也不应作为评估研究影响

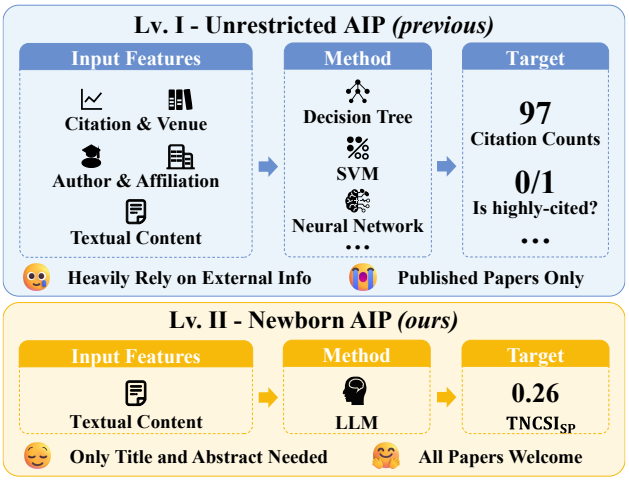

图 1: 文章影响力预测（AIP）分类法：由于几乎没有其他 Lv.II 方法，“新生 AIP”部分代表了所提出的方法，它以“双盲同行评审”的方式预测未来的学术影响力。

力的唯一指标。例如，在蓬勃发展的人工智能领域，一篇论文的引用次数达到 100 次可能不足为奇，但在古生物学等相对狭窄但同样重要的领域，一篇论文的引用次数达到 100 次 (Turner 2011) 则可被视为基石。由于自动化科研系统大多根据引用次数估算价值，这种局限性无疑会削弱其从其他领域收集知识的能力，从而降低知识合成的效率。

为了解决将引用次数作为回归目标的潜在问题，我们首先从话题归一化引用超越指数（TNCSI）的设计原则中汲取灵感 (Zhao et al. 2024)，并对其进行了有针对性的改进，使其适用于预测不同领域新生论文的影响力。改进后的指标被命名为 TNCSI_{SP}，其中 SP 代表同一时期，以强调所提出的指标能够比较不同时间段的论文。由于贡献、新颖性和见解等关键要素通常反映在标题和摘要中，因此我们认为文章的“价值”通常可以通过“文本”来评估。因此，我们尝试通过只输入标题和摘要来回归 TNCSI_{SP}，以微调大语言模型，从而获得可靠的影响力预测。

为了更清楚地了解影响力预测方法的现状，我们总结并介绍了基于预测所需信息的分类法（见图 1）。第一层被称为无限制文章影响力预测。在这种情况下，预测可以依赖于外部历史信息和作者的声誉。这是目前大多数方法所处的水平。第二个层次被称为新生文章影响力预测，它特别强调只根据文章本身来预测影响力。这项任务类似于双盲审查过程，即模型在没有任何作者和所属单位信息、出版细节或早期引用数据的情况下预测未来的影响力。这种方法对于筛选新上传的稿件（如 arXiv 预印本和会议论文）尤其有价值，因为它可以帮助研究人员有效地识别出最有前途的文章。在本文中，我们将重点关注最具挑战性但最有价值的任务：新生文章影响力预测。

总之，这项工作的核心贡献如下：

- **新的任务**: 我们引入了一个分类法，并定义了一项名为“新生文章影响力预测”的新任务，其目的是在没有外部信息的情况下准确预测新发表文章的学术影响力。
- **新的方法**: 我们对 TNCSI 进行了有针对性的改进，并首次证明了微调 LLM 能够在”双盲评论”设置中预测新生文章的未来影响力。
- **新的数据集**: 因此，我们构建并发布了 TKPD 和 NAID 数据集。这两个数据集分别用于指导 ChatGPT 生成主题关键短语，以及训练最先进的 LLM，以准确预测文章影响力。
- **应用**: 最后，我们讨论并举例说明了所提方法在现实世界中的应用，特别是在预测 2024 年发表的期刊论文的影响力方面，希望能激励更广泛的研究领域取得进一步进展。

Related Work

文献计量学是一个利用定量分析和统计方法评估学术出版物影响力的研究领域。通常，文献计量学可分为两大类：期刊评价指标和单篇文章评价指标。正如《莱顿宣言》（Leiden Manifesto）(Hicks et al. 2015) 和《DORA 宣言》(DORA Declaration) (San Francisco 2018) 所建议的，不要使用基于期刊的指标来衡量单篇研究文章的质量。因此，在本文中，我们不打算使用任何期刊级别的文献计量指标（如 JIF）(Garfield 1955) 作为输入或预测目标。相反，我们专注于单篇文章的文献计量指标。表 1 说明了它们之间的差异。虽然 FWCI 和 RCR 都是很好的度量指标，但它们的非规范化数值特性可能会

影响神经网络的收敛性。TNCSI (Zhao et al. 2024) 由 Zhao 等人提出，具有明确的物理意义和良好的数学特性，代表了一篇文章的影响力超过同一领域其他文章的概率（范围在 0 到 1 之间）。然而，TNCSI 最初是为评估不同领域的评论性论文而设计的，因此并不适合评估普通的研究论文。此外，TNCSI 主要侧重于评估综述论文的现有影响力，并没有对不同年份发表的论文的影响力进行归一化处理。这可能会导致新生文章影响力预测任务中潜在的不公平比较。因此，针对 TNCSI 的局限性，我们提出了一个改进版本。更多细节请参见方法部分。

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

表 1: 几种用于评估学者影响力的文章级别计量指标：Value、Field 和 Time，分别表示该指标是否为 0 到 1 之间的值，是否允许跨领域比较，以及是否适合用于比较不同时间发布的论文。这些归一化操作有助于网络训练。

文章影响力预测方法通常采用机器学习方法来预测文章的未来影响力。大多数现有方法倾向于利用文章统计特征、作者特征、期刊属性和历史引用数据来辅助决策树、LSTM、MLP 和其他机器学习算法进行预测 (Fu and Aliferis 2008; Wang, Yu, and Yu 2011; Qiu and Han 2024; Kousha and Thelwall 2024)。Ruan 等人 (Ruan et al. 2020) 利用四层反向传播（BP）神经网络，利用论文、期刊、作者、参考文献和早期引文的多个相关特征，提高了对五年引文数的预测精度。Ma 等人 (Ma et al. 2021) 提出了一种引文计数预测模型，该模型使用早期引文和论文序列特征作为输入，并采用 Bi-LSTM 进行最终预预测。另一种著名的基于引文的机器学习方法利用静态特征和随时间变化的引文特征来预测潜在的优秀论文 (Hu, Cui, and Lin 2023)。在 ABBAS 的研究 (Abbas et al. 2023) 中，提出了一种基于 MLP 的方法，只利用外部特征来预测未来的引文数量，取得了不错的效果，NDCG 达到 0.95。Zhang 等人 (Zhang and Wu 2024) 发现，通过利用早期引文数据，对不同领域的论文采用不同的模型可显著提高预测的准确性。De (de Winter 2024) 尝试指导 ChatGPT-4 从多个角度对 2000 多篇论文摘要进行评分，发现评分与 Mendeley 读者数量的斯皮尔曼相关系数大于 0.4，与引用次数的相关系数为 0.18。据我们所知，目前还没有一种方法能够仅根据文章的内部内容准确预测其影响力。

大语言模型在过去几年中展现出了强大的长文本建模能力，并被广泛应用于各种 NLP 任务中，包括对话系统、机器翻译、情感分析等 (Zhao et al. 2023; Tu et al. 2024; Jiang et al. 2024)。许多商业化的大型语言模型 (OpenAI 2022, 2023; Google 2024; Kimi.ai 2024) 都无法公开访问，这使得我们无法对其进行微调或指令调整。因此，我们把目光转向了几个优秀的开放式大型语言模型。LLaMA series (Touvron et al. 2023; AI 2024) 是由 Meta AI 创建的高级语言模型，有 7B 到 70B 参数的多个版本。它在大多数任务中都表现出不俗的性能，已被广泛应用于各种应用中。除了 LLaMA 之外，还有其他一些著名的开源大型语言模型，如 Qwen (Bai et al. 2023)、Mistral (Jiang et al. 2023)、Falcon (Almazrouei et al. 2023) 等。无论具体的大语言模型是什么，它们最初都是为自回归文本生成而开发的。在本文中，我们仅使用第一个生成的标记进行数值回归。我们将在方法和实验部分对这些模型进行详细说明和综合评估。

Approach

对 TNCSI 进行针对性改进

如相关工作部分所述，TNCSI 存在一定的局限性，例如仅限于评估综述论文，并且只考虑了文章的累积影响力。我们对其计算过程进行了详细分析，并找出了这些局限性背后的原因。

首先，TNCSI 需要一个预定义的提示模板，以指导 ChatGPT 根据给定的标题和摘要生成相应的综述研究领域。原始提示是专门为综述论文而非普通研究论文设计的。因此，在普通论文上直接使用他

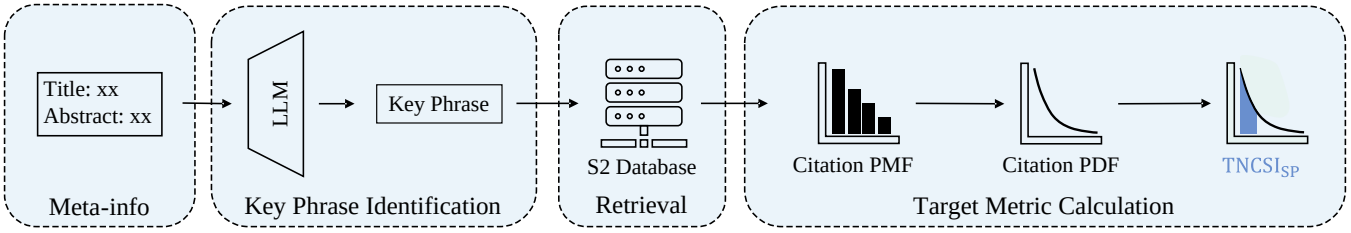

图 2: 计算 TNCSI$_{SP}$ 的流程图：TNCSI$_{SP} \in [0,1]$ 表示某篇论文的引用次数超过同一领域和时间段内其他论文的概率。“S2” 指的是 Semantic Scholar。

们的提示会导致性能不佳。其次，TNCSI 主要考虑文章发表后的累积影响力。然而，在构建文章影响力预测任务的数据集时，这种方法可能会导致潜在的不公平比较问题。具体来说，较早发表的论文的 TNCSI 值通常高于近期发表的论文。这可能会混淆网络的学习过程，使 LLM 难以建立文本特征与其影响力值之间关系的模型。

基于上述讨论，我们对 TNCSI 进行了有针对性的改进，并将改进后的度量命名为 TNCSI$_{SP}$。与 TNCSI 的计算过程类似，所提出的 TNCSI$_{SP}$ 的计算过程也分为三个步骤。第一步，利用精心设计的提示来引导 ChatGPT（目前称为 gpt-3.5-turbo-0125）识别文章的主题关键词。我们设计并测试了多种提示模板，以便从不同角度识别文章关键词。为了进一步减少个人认知偏差，我们在提示创建过程中寻求了众多研究人员的帮助。所有提示模板都在人工标注的数据集上进行了测试，以评估其在关键短语识别任务中的相应性能。第二步是使用 ChatGPT 生成的关键短语从语义学者 API 中检索 1000 篇相关论文及其相关信息（如引用次数）。TNCSI 考虑的是整个时间范围内的引用计数，而 TNCSI$_{SP}$ 则不同，它关注的是发表日期前后 6 个月内的同期论文。

这种方法确保了每篇论文只与在相似时间内发表的其他论文进行比较，从而最大限度地减少了老论文因发表时间较长而积累的引文优势。因此，这种方法使 TNCSI$_{SP}$ 具有对不同发表时间的引文影响力进行归一化的能力。最后一步与 TNCSI 保持一致。简化的数学表达式如下：

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

此处，$C = 1000$ 指检索到的论文总数。Count(p_x) 表示有 x 引用的论文 p 的数量。$P(X = x)$ 是一个离散概率分布，描述了在检索到的 C 论文中，某篇论文被引用次数正好为 x 的概率。

在他们的工作中 (Zhao et al. 2024)，$P(X = x)$ 已被深入讨论，并被证明遵循指数衰减分布。因此，可以使用最大似然估计法将其转换为概率密度函数。如公式 (2) 所示，我们可以通过计算相应定积分的值，推导出最终的 TNCSI$_{SP} \in [0,1]$ 。

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

其中，$cites$ 表示被评估论文的引用次数。

用于预测新生文章影响力的大语言模型

大型语言模型的自回归机制已经得到了充分的论证 (Zhao et al. 2023)。从本质上讲，这些纯解码器模型以顺序的方式生成文本，每个标记的预测都依赖于前一个标记所提供的上下文。这种范式使其能够充分利用无标记数据进行自我监督学习。

在本文中，我们保持大语言模型的自回归生成方案不变。不过与传统的文本生成不同，我们仅关注模型在响应用户输入时自回归生成的第一个令牌。具体来说，假设当前的输入序列为 $\{w_1, w_2, \ldots, w_t\}$。LLM 与下一个令牌 w_{t+1} 的生成之间的关系可以表示为：

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

其中 LLM$(\cdot)$ 表示一个大型语言模型，它可以根据输入标记序列 $\{w_1, w_2, \ldots, w_t\}$ 预测序列中的下一个标记。

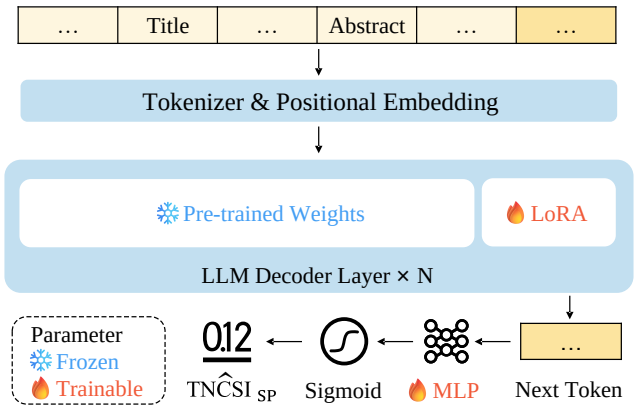

图 3: LLM 作为学术影响力预测器：所提方法的整体框架。仅使用下一个令牌（第一个生成的令牌）来回归 TNCSI_{SP}。

为了促进 LLM 对单个数值的预测，我们采用一个简单的多层感知机（MLP）将 $w_{t+1} \in \mathbb{R}^{B\times 1\times D}$ 转换为一个实数 $v \in \mathbb{R}$。然后，该值 v 被输入到 Sigmoid 函数中，从而得到预测值 $\widehat{\text{TNCSI}}_{\text{SP}} \in [0,1]$。其中，$B$ 表示批量大小，D 是维度，$\mathbb{R}$ 表示实数集。该过程可表示为以下方程：

$$\hat{TNCSI}_{SP} = \sigma(\text{MLP}(w_{t+1})), \tag{4}$$

其中，$\widehat{\text{TNCSI}}_{\text{SP}}$ 是通过将 w_{t+1} 传递通过一个 MLP 后，再通过一个 Sigmoid 函数 σ 计算得到的。

最后，我们的目标是最小化均方误差（MSE）损失，以使得预测输出 $\widehat{\text{TNCSI}}_{\text{SP}}$ 与通过先前统计计算得到的 TNCSI_{SP} 对齐。

在实际应用中，大语言模型中庞大的参数量需要大量的计算资源进行训练，这超出了我们的实际能力。因此，我们采用了低秩矩阵分解（LoRA）(Hu et al. 2021) 和模型量化技术 (Dettmers et al. 2022) 来降低计算资源消耗，并加速网络的训练和推理过程。关于更多细节，我们建议读者参考原始论文。

数据集构建

我们共构建了两个数据集，分别是主题关键词数据集（TKPD）和标准化文章影响力数据集（NAID）。这两个数据集各自有不同的用途，下面将更详细地描述。

主题关键词数据集: TKPD 包含 251 条条目，涵盖了来自人工智能各个领域的随机文章的标题、摘要以及核心任务或领域名称。为了减轻研究中的主观性，并确保注释的一致性，我们采用了一名经验丰富的 AI 研究员进行关键短语的人工标注，并邀请了另外三位研究员对标注结果进行复核。由于数据标注需要专业知识，本文无法标注非 AI 领域的文章。不过，我们认为人工智能领域内不同子领域之间的不一致性足以模拟不同学科之间的差异。

标准化文章影响力数据集: NAID 用于训练大语言模型预测文章的影响力。它包括标题、摘要以及相应的 TNCSI_{SP} 等数据。NAID 包含来自各个 AI 领域的超过 12,000 条数据条目，不包括综述论文，涵盖了在 2020 到 2022 年期间上传至 arXiv 的“cs.CV”、“cs.CL”和“cs.AI”类别的论文。特别地，“cs.AI”类别涵盖了数学、物理学和认知科学等多个学科，从而将训练数据扩展到 AI 领域之外。NAID 是一个均匀分布的数据集，这意味着论文的来源、原始发表年份以及相应的 TNCSI_{SP} 值都呈均匀分布。

Experiments

指标

平均绝对误差: MAE 用于评估预测精度。它是用来衡量预测值与真实值 y_i 之间差异的指标，定义如下：

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

其中，n 表示测试集中的样本数量，y_i 表示实际输出（如 TNCSI_{SP}），$\hat{y}_i$ 表示预测值（如 $\hat{TNCSI}_{SP}$），$|y_i - \hat{y}_i|$ 是实际值与预测值之间的绝对差异。通常较低的 MAE 表示模型预测的准确性较高。

标准化折扣累积增益 (Järvelin and Kekäläinen 2000): NDCG 是另一个用于评估预测效果的指标。NDCG 最初是为推荐系统开发的，用于衡量文档在推荐列表中的位置所带来的增益，计算方法如下：

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

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**

表 2: 与以往方法的比较：所提出的方法在各方面表现出显著的优势。向上的箭头表示更高的值更好，反之亦然。粗体字表示在所有方法中表现最好的。Ori. Lv. 代指原始研究的分类层级，Target 表示对应研究中预测的目标类型。有关更多复现细节，请参见附录。

其中，$\text{DCG@K} = \sum_{i=1}^{K}(2^{\hat{y}_i} - 1)/\log_2(i+1)$，而 $\text{IDCG@K} = \sum_{i=1}^{K}(2^{y_i} - 1)/\log_2(i+1)$。$K = 20$ 表示推荐列表的排名截断位置。NDCG 是一个范围从 0 到 1 的指标，得分越接近 1 表示更有影响力的文档排名越靠前，从而体现了更好的性能。

标准化编辑距离 (Yujian and Bo 2007): NED 是一个度量两个字符串相似度的指标，它通过将编辑距离标准化为较长字符串的长度来计算。其定义如下：

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

其中，$\text{ED}(A, B)$ 是字符串 A 和 B 之间的编辑距离，$\max(|A|, |B|)$ 是较长字符串的长度。NED 值越低，表示两个字符串越相似。

与以往方法的比较

我们使用 NDCG 来评估不同方法在识别高影响力论文方面的有效性，考虑到它们的不同预测目标。为了确保公平性，我们排除了先前方法依赖的外部数据。复现细节已在附录中提供。

表 2 清晰地展示了我们提出的方法与先前方法在识别潜在高影响力论文方面的性能差异。在新生文章影响力预测的场景中，所提出的方法相较于早期的代表性工作表现出显著的优势。大多数一级方法在没有外部信息的情况下表现不佳。例如，基于 LSTM 的方法 (Ma et al. 2021) 在利用外部信息时报告的 NDCG 为 0.84，但当仅依赖标题和摘要时，其性能显著下降至 0.196，表明其在有效地将语义特征映射到目标 TNCSI_{SP} 上的能力有限。ChatGPT 生成和 LLaMA-3 生成方法在识别高影响力论文方面的较差表现表明，零样本 LLM 生成方法仍需进一步探索。总之，我们认为，所提出方法的显著性能提升可以归功于 LLaMA-3 的广泛基础知识以及在微调过程中纳入了 TNCSI_{SP} 指标，这增强了其在各个领域和时间段内识别有影响力语义特征的能力。

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

表 3: 不同的大语言模型在 NAID 测试集上的性能比较：Memory 表示推理过程中最小的内存使用量。

各种大语言模型的表现

作为本文的核心任务，我们全面评估了不同大语言模型在 NAID 测试集上的表现。如表 3 所示，LLaMA-3-8B 实现了最佳的整体性能。有趣的是，我们观察到 MAE 和 NDCG 并不总是呈反向关联；例如，尽管 Falcon 的 MAE 低于 Phi-3，但其 NDCG

稍低。这表明，Falcon 在预测低影响力论文方面更为准确，但在高影响力论文的预测中效果较差。由于我们的主要关注点是识别高影响力论文，在这种情况下，较高的 NDCG 通常比较低的 MAE 更具优势。

Qwen 系列被选中进一步探讨模型大小对文章影响力预测任务性能的影响。与 LLaMA 系列相比，Qwen 系列具有更多官方模型，且其参数规模较小，具体为 0.5B、1.5B 和 7B。我们在 NAID 训练集上训练了这些模型，每个模型的测试结果如图 4 所示。可以观察到，随着模型参数规模的增大，性能相应地得到了提升。

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**

表 4: 用于识别主题关键词的各种用户提示比较。

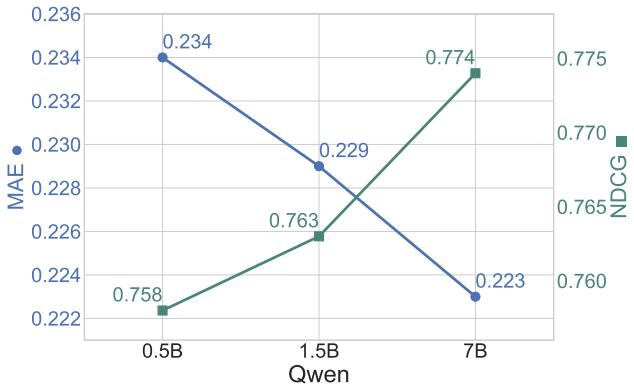

图 4: 各种模型参数对性能的影响：模型参数数量越大，性能越好。

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**

表 5: 引导大语言模型预测未来影响力的各提示比较。

提示工程的效果

本文对两个任务进行了提示工程：一是识别用于计算 TNCSI_{SP} 的主题关键词，二是引导大语言模型进行预测。

用于识别关键词: 我们进行了大量实验，测试不同模型和提示在 TKPD 上的表现。表 4 报告了 3 个代表性提示在 TKPD 上的 NED。最终，我们采用了最后一行的提示模板，并结合 gpt-3.5-turbo-0125 来生成主题关键词。扩展的实验记录可在补充材料中找到。

用于引导大语言模型: 如表 5 所示，我们测试了几种提示模板，将标题和摘要包裹后输入到微调后的 LLM 中。尽管使用了 PEFT，提示模板的变化仍然会影响性能；更详细的描述通常会带来更好的结果。然而，过于详细的提示也可能导致 NDCG 轻微下降。

TNCSI_{SP} 的比较分析

我们已经在来自不同年份的文章以及各种回归目标上训练了 LLaMA-3，以展示所提 TNCSI_{SP} 的优越性。如图 5 所示，当目标是改进后的 TNCSI_{SP} 时，模型在针对不同年份的文章时提供了更稳定的预测。表 6 进一步展示了 TNCSI_{SP} 的泛化能力。它使得各种类型的模型能够更好地抵抗随时间积累的偏差。这表明，TNCSI_{SP} 使模型能够识别不同年份

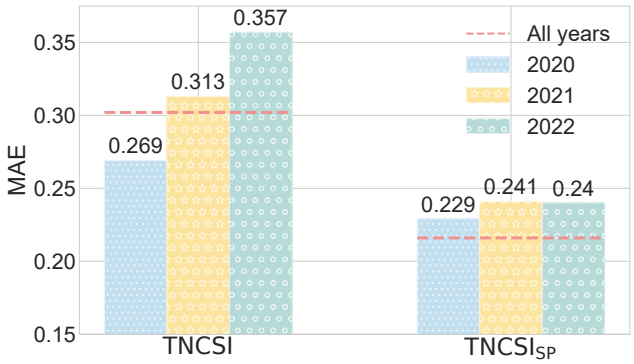

图 5: 不同预测目标对性能的影响：TNCSI$_{SP}$ 在使用来自不同年份的训练数据时，表现优于 TNCSI。

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**

表 6: 使用 TNCSI$_{SP}$ 指标的性能比较：所有方法以 TNCSI$_{SP}$ 为目标时都表现出提升。输入与表 2 一致。

中高影响力文章共享的语义特征，从而在整体任务性能上取得显著提升。

Applications

在本节中，我们展示了一个有趣的例子——期刊平均影响力预测，以进一步证明我们方法在实际应用中的有效性。

从理论上讲，不同分区的期刊预计会表现出不同的平均影响力。因此，我们引导 LLaMA-3 预测 2024 年发表的计算机科学领域多个期刊中文章的平均 TNCSI$_{SP}$，这些期刊来自不同的 JCR 分区。由于 LLaMA-3 的训练数据仅延续到 2023 年初，因此该模型很可能未曾遇到这些文章，从而大大降低了数据泄露的风险。值得注意的是，期刊的影响因子通常会受到少数高度引用论文的显著影响 (Lei and Sun 2020; Leydesdorff 2012)。为此，我们分析了从不同分区的多个期刊中随机选取的 500 多篇文章的影响

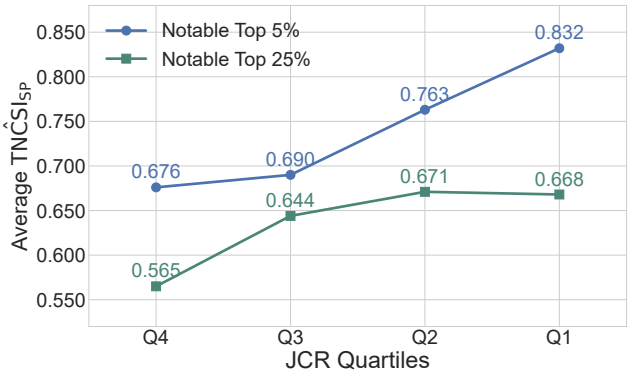

图 6: 不同 JCR 分区期刊的预测 TNCSI$_{SP}$ 值：较高的分区显示较高的预测值。为了避免潜在的利益冲突，我们用 Q1、Q2、Q3 和 Q4 分别表示来自 JCR 分区 1、2、3 和 4 的期刊文章。

力预测，重点关注每个分区中排名前 5% 和 25% 的显著论文的平均预测 TNCSI$_{SP}$。在图 6 中，我们观察到显著的前 5% 文章的预测影响力与其所属分区之间存在明显的正相关性。尽管 Q2 分区中排名前 25% 的文章的预测影响力略高于 Q1，但这仍被视为一种合理现象。

除了期刊影响力预测，我们的系统还在其他多种实际应用中具有潜力。例如，考虑到每天都有大量新提交的预印本论文，所提出的方法也可以帮助高效识别值得进一步审查的高质量研究。它可以显著减少研究人员在审阅大量 arXiv 论文时所花费的时间，从而提高整体研究效率

Conclusion

在本文中，我们展示了大语言模型在仅凭标题和摘要预测新生论文的量身定制 TNCSI$_{SP}$ 的潜力。构建并使用了包含超过 12,000 条记录的 NAID 数据集，用于微调各种先进的 LLM 模型。实证评估表明，LLaMA-3 模型在仅依赖内部信息的情况下，具有 0.216 的平均绝对误差（MAE）和 0.901 的 NDCG@20，显著超过了之前方法的表现。此外，我们的方法预测的影响力值与 2024 年发表文章的期刊分区排名之间存在强烈的正相关性，展示了我们方法在实际应用中的可行性。总体而言，所提方法有效地从 0 到 1 估计了新发布论文的未来影响力评分，

为个人、机构和自动化科研系统带来了显著的益处。

Ethical Statement

我们意识到过度优化标题和摘要可能会造成操纵。研究人员必须避免过度美化标题和摘要，特别是通过虚假宣称尚未实现的性能或过度夸大方法的重要性，以试图操控预测的影响力值。

由于诸如 Semantic Scholar API 访问频率限制等因素的约束，我们无法构建更大规模的数据集。因此，我们提出的方法仅作为初步的探索性方法。该方法生成的预测是概率估计，绝不应被视为对文章质量的最终评估。该方法旨在提供额外的见解，不能替代现有的同行评审过程，同行评审仍然是维护学术研究完整性和严谨性的关键。作者对基于预测结果做出的任何决策不承担责任。

Acknowledgements

这项研究得到了中国国家自然科学基金青年科学基金（资助号：62206134）、中央高校基础研究基金（070-63233084）以及天津视觉计算与智能感知重点实验室（VCIP）的支持。计算工作得到了南开大学超级计算中心（NKSC）的支持。本工作还得到了中国国家科学基金（资助号：62361166670）的资助。

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.

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*.