SciImpact: A Multi-Dimensional, Multi-Field Benchmark for Scientific Impact Prediction

Anonymous ACL submission

Abstract

The rapid growth of scientific literature calls for automated methods to assess and predict research impact. Prior work has largely focused on citation-based metrics, leaving limited evaluation of models' capability to reason about other impact dimensions. To this end, we introduce SciImpact, a large-scale, multi-dimensional benchmark for scientific impact prediction spanning 19 fields. SciImpact captures various forms of scientific influence, ranging from citation counts to award recognition, media attention, patent reference, and artifact adoption, by integrating heterogeneous data sources and targeted web crawling. It comprises 215,928 contrastive paper pairs reflecting meaningful impact differences in both short- (e.g., Best Paper Award) and long-term settings (e.g., Nobel Prize). We evaluate 11 widely used large language models (LLMs) on SciImpact. Results show that off-the-shelf models show substantial variability across dimensions and fields, while multi-task supervised fine-tuning consistently enables smaller LLMs (e.g., 4B) to markedly outperform much larger models (e.g., 30B) and surpass powerful closed-source LLMs (e.g., `o4-mini`). These results establish SciImpact as a challenging benchmark and demonstrate its value for multi-dimensional, multi-field scientific impact prediction. Our benchmark and code are available at `https://gitlab.com/user-paper-review/SciImpact.git`.

1 Introduction

As scientific literature continues to grow exponentially (Dong et al., 2017), researchers face unprecedented challenges in identifying influential research from an ever-expanding body of work. This challenge motivates the development of techniques for predicting which studies are likely to become influential in the future (Xia et al., 2023), thereby supporting effective knowledge acquisition, scientific evaluation, and decision-making. Prior work on scientific impact prediction has largely focused on citation count prediction and its variants (Dong et al., 2015; Li et al., 2019b; Hirako et al., 2023). However, while citations are positively correlated with some other measures of scientific recognition (Jin et al., 2021; Zhang, 2025), they alone are insufficient to capture the full range of factors that reflect impact (Radicchi et al., 2017). In particular, the following aspects also warrant consideration.

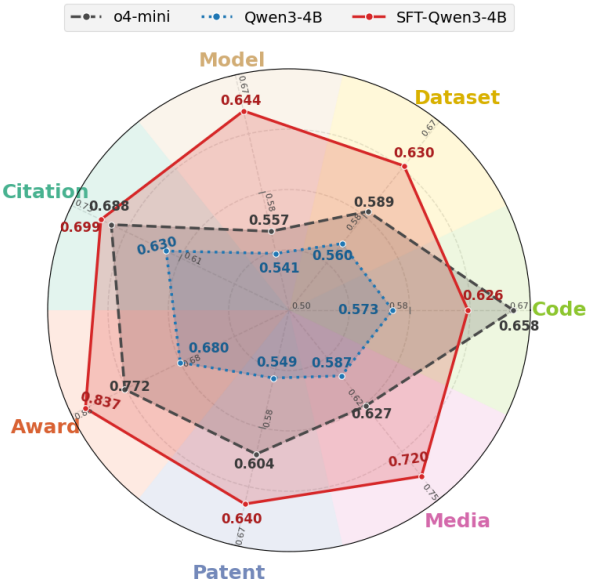

Figure 1: Performance of `o4-mini`, off-the-shelf `Qwen3-4B`, and supervised fine-tuned `Qwen3-4B` across the seven impact dimensions on SciImpact. Supervised fine-tuning (SFT) substantially enhances a 4B open-weight model's ability to predict scientific impact across all dimensions, enabling it to rival or surpass stronger closed-source models.

Award Recognition. Prize-winning topics produce 47% more star scientists and attract 37% more new entrants (Jin et al., 2021). In physics, chemistry, medicine, and economics, predicting which papers may lead their authors to win a Nobel Prize is an annually high-profile task closely related to impact prediction. In computer science conferences, best paper award prediction offers another perspective

on academic impact (Huang, 2023).

Public Use. Scientific articles are not only cited within the "ivory tower" of academia but are also consumed in public domains, such as technological outlets (e.g., patents) and societal channels (e.g., news and social media). Previous studies (Yin et al., 2022; Zhang, 2025) have shown that papers referenced in patents or media posts are 5 to 18 times as likely to become high-impact compared to a randomly selected paper.

Artifact Adoption. Scientific papers, especially in computer science, are often accompanied by artifacts such as codebases (Papers with Code, 2019), constructed datasets (Yang et al., 2024b), and pretrained models (Liang et al., 2024) hosted on platforms like GitHub or Hugging Face. Intuitively, the number of times these byproducts are downloaded or starred by users on such platforms also serves as a crucial measure of a paper's impact.

To bridge the gap between prior work predominantly targeting citation count prediction and the multi-faceted impact criteria outlined above, in this paper, we propose SCIIMPACT, a comprehensive, multi-dimensional, and multi-field benchmark for scientific impact evaluation. As shown in Table 1, SCIIMPACT covers 7 distinct impact dimensions (Citation, Award, Patent, Media, Code, Dataset, and Model), strictly more than any single existing data source to the best of our knowledge. Moreover, SCIIMPACT goes beyond computer science and biomedicine papers emphasized in previous scientific literature understanding studies, encompassing papers from natural sciences, engineering, social sciences, and humanities (corresponding to all 19 fields in the Microsoft Academic Graph (Shen et al., 2018)). This results in 215,928 contrastive paper pairs, enabling models to predict which paper in each pair has greater impact in a given dimension. It is worth noting that, in constructing this benchmark, we not only curate data from fragmented and heterogeneous existing resources but also crawl missing data for specific dimensions and fields from the web (e.g., MDPI Best Paper Awards and GitHub forks).

Based on SCIIMPACT, we conduct a comprehensive evaluation of 11 prominent large language models (LLMs) for scientific impact prediction, including 3 closed-source models and 8 open-source models. In addition, we aggregate training data across all impact dimensions and perform multi-task instruction tuning to train unified scientific impact prediction models using `Qwen3-4B` (Yang et al., 2025) and `Llama-3.2-3B` (Grattafiori et al., 2024) as backbones. Figure 1 compares a representative closed-source model (`o4-mini`), an off-the-shelf open-weight model (`Qwen3-4B`), and its supervised fine-tuned counterpart (`SFT-Qwen3-4B`) across the seven impact dimensions, illustrating the benefits of fine-tuning on SCIIMPACT. (We provide the corresponding comparison across fields for the same three models in Appendix A.) Overall, our results show that the fine-tuned 4B model delivers the strongest average performance and is competitive with leading closed-source baselines, underscoring the value of SCIIMPACT for scientific impact prediction as a multi-dimensional task.

The contributions of our work are as follows:

- We broaden the scope of scientific impact prediction by framing impact as a multi-dimensional concept that goes beyond citation counts, incorporating diverse forms of award recognition, public use, and artifact adoption.
- To support this perspective, we introduce SCIIMPACT, a comprehensive, multi-dimensional, multi-field benchmark for scientific impact evaluation, built via curated integration of fragmented, heterogeneous resources and targeted web crawling to fill missing dimensions and fields.
- We conduct large-scale experiments on SCIIMPACT to evaluate various prominent LLMs and train unified scientific impact prediction models via multi-task instruction tuning. Results show that this fine-tuning enables relatively small LLMs to achieve superior performance compared to much larger or stronger models.

2 Related Work

2.1 Evolution of Scientific Impact Prediction

Quantitative studies of scientific literature primarily use citation counts as a proxy for impact (Wang et al., 2013; Sinatra et al., 2016). Early work predicts future citations from features available at or shortly after publication, such as author history and bibliometric cues (Castillo et al., 2007; Fu and Aliferis, 2008; Ibáñez et al., 2009). Later studies reveal that heterogeneous citation trajectories, motivating dynamic models that account for temporal effects including aging and cumulative advantage (Chakraborty et al., 2014; Xiao et al., 2016). With the advent of deep learning, sequence-based approaches further improve citation forecasting by modeling temporal dependencies (Yuan et al.,

	Dimension Coverage							Field Coverage		
	Citation	Award	Patent	Media	Code	Dataset	Model	Comp. Sci.	Biomedicine	Other Fields
Li et al. (2019a)	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓
Li et al. (2019b)	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
Hirako et al. (2023)	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗
Lin et al. (2023)	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓
Yang et al. (2024b)	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗
Liang et al. (2024)	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
Zhang (2025)	✓	✗	✓	✓	✓	✗	✗	✓	✗	✗
SciImpact (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between SciImpact and existing data sources.

2018), and auxiliary signals such as peer review text enhance prediction (Li et al., 2019b). More recently, LLMs have been applied to citation prediction tasks (Zhao et al., 2025; Lu et al., 2025).

Beyond citations, impact includes artifact adoption and external attention. Work analyzes popularity signals in open-source ecosystems, such as GitHub stars and their relationship to downstream usage (Ren et al., 2020; Koch et al., 2024), as well as dataset and model reuse on platforms like Hugging Face (Koch et al., 2021; Yang et al., 2024b; Liang et al., 2024). External influence is also quantified using patents, media, and policy documents alongside scholarly citations (Yin et al., 2022; Zhang, 2025).

Overall, existing studies typically focus on a single proxy, platform, or prediction horizon, and lack a unified benchmark for systematic comparison across fields and impact dimensions. These limitations motivate SciImpact, which provides standardized evaluation over diverse disciplines and heterogeneous indicators of scientific influence.

2.2 Datasets for Science Literature Analysis

Advances in science literature analysis are enabled by large-scale scholarly datasets. Early work widely relies on the Microsoft Academic Graph (MAG; Shen et al., 2018), which also supports curated resources such as Nobel-laureate publication datasets (Li et al., 2019a). Following MAG's discontinuation, OpenAlex (Priem et al., 2022) emerges as a fully open alternative with broad metadata and citation coverage. Complementary resources improve cross-disciplinary analysis, including MAPLE for field-aware topic tagging (Zhang et al., 2023) and SciSciNet as an integrated data lake linking publications to external signals (Lin et al., 2023). Recent datasets are often released with task-specific benchmarks, such as impact forecasting on evolving scholarly graphs (Gu and Krenn, 2024), interdisciplinary link prediction (Rezaee et al., 2025), and text impact prediction for newborn papers using LLMs (Zhao et al., 2025). While these resources advance meta-scientific research, they typically center on a single task or dataset family. SciImpact complements them by providing a unified benchmark for scientific impact prediction across multiple fields, dimensions, and time horizons.

3 SciImpact Benchmark

We now describe the construction of our SciImpact benchmark. Each instance in SciImpact is a *contrastive* pair of artifacts $(\mathcal{A}^+, \mathcal{A}^-)$, where $\mathcal{A}^+$ exhibits a higher impact signal than $\mathcal{A}^-$ within a certain dimension. Here, "artifacts" may refer to research papers or their associated model cards, dataset cards, or repository README files. SciImpact covers 19 fields, spanning art, biology, business, chemistry, computer science, economics, engineering, environmental science, geography, geology, history, materials science, mathematics, medicine, philosophy, physics, political science, psychology, and sociology. We construct the benchmark by integrating online resources with existing datasets through a three-stage pipeline: (1) candidate retrieval, (2) impact labeling and pair generation, and (3) filtering and quality control. Figure 2 summarizes the pipeline.

For the impact metric $y(\mathcal{A})$, we consider seven dimensions that capture both academic and broader forms of influence: (1) citations; (2) award recognition (including Best Paper Awards from major computer science conferences, the Nobel Prize for physics/chemistry/medicine, and MDPI Best Paper Awards for other fields); (3) patent references; (4) media attention (combining news coverage and social media mentions); (5) GitHub stars; (6) Hugging Face dataset downloads; and (7) Hugging Face model downloads. Table 2 summarizes the thresholds applied to construct contrastive pairs for each impact dimension.

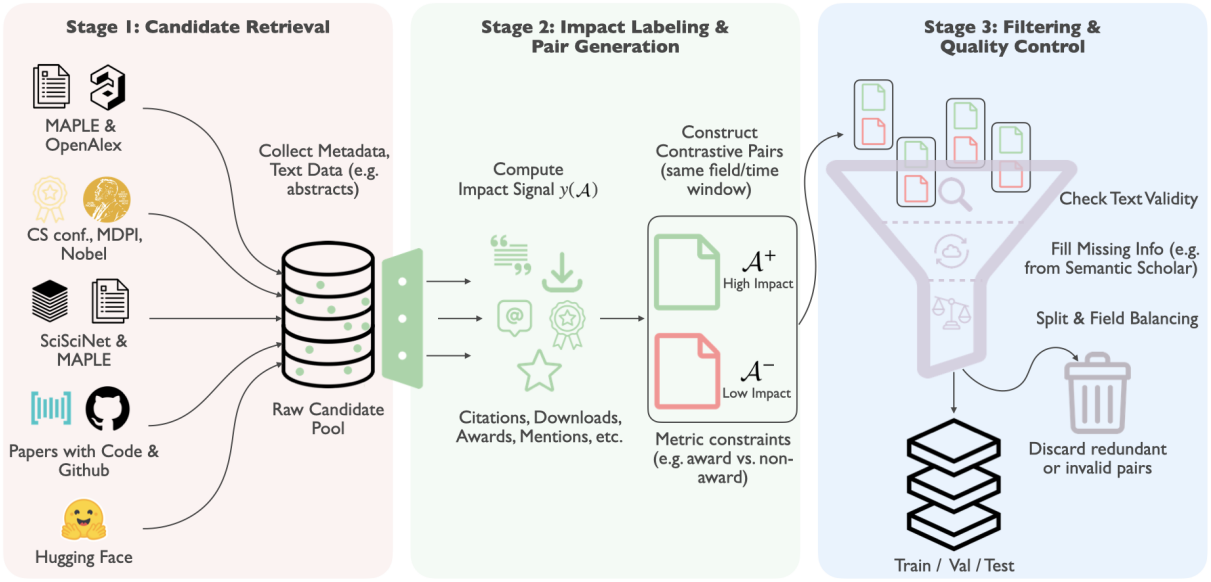

Figure 2: Overview of the SCIIMPACT benchmark curation pipeline, including candidate retrieval, impact labeling and pair generation, and filtering and quality control.

Dimension	Pair Construction Rule
Citation	$y(\mathcal{A}^+) \geq 10,\ y(\mathcal{A}^-) \geq 10,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$
Award	$y(\mathcal{A}^+) = \texttt{True},\ y(\mathcal{A}^-) = \texttt{False}$
Patent	$y(\mathcal{A}^+) \geq 5,\ y(\mathcal{A}^-) \geq 5,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$
Media	$y(\mathcal{A}^+) \geq 5,\ y(\mathcal{A}^-) \geq 5,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$
Code	$y(\mathcal{A}^+) \geq 10,\ y(\mathcal{A}^-) \geq 10,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$
Dataset	$y(\mathcal{A}^+) \geq 10,\ y(\mathcal{A}^-) \geq 10,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$
Model	$y(\mathcal{A}^+) \geq 10,\ y(\mathcal{A}^-) \geq 10,\ \frac{y(\mathcal{A}^+)}{y(\mathcal{A}^-)} \geq 2$

Table 2: Impact dimensions and thresholding rules used to construct contrastive pairs in SCIIMPACT. **Note:** For award recognition, $y(\mathcal{A})$ is a boolean indicator reflecting whether the artifact receives the corresponding award. For all other dimensions, $y(\mathcal{A})$ is a nonnegative count (e.g., citation count).

3.1 Stage 1: Candidate Retrieval

Citation. We first retrieve candidate papers from MAPLE (Zhang et al., 2023), which collects research articles published in the top-100 venues of each of the 19 fields. Publication years are restricted to 2001-2020 to allow sufficient time for citations to accumulate. We then match MAPLE entries with OpenAlex (Priem et al., 2022) to obtain the title, abstract, year, and citation count up to mid-2025, which serves as $y(\mathcal{A})$. To ensure a fair comparison, $\mathcal{A}^+$ and $\mathcal{A}^-$ in a contrastive pair are required to be published in the same year. Note that the Citation dimension encompass scientific impact prediction over different time horizons: pairs published in 2001 correspond to longer-term impact prediction, while pairs from 2020 represent a shorter-term prediction horizon.

Award. We crawl award data from three sources depending on the field: (1) Best Paper Awards from major computer science conferences (Huang, 2023), (2) Nobel Prize-winning papers for physics, chemistry, and medicine (Li et al., 2019a), and (3) MDPI Best Paper Awards for the remaining fields (MDPI, 2025). We link each award-winning paper to OpenAlex via DOI matching and collect the required bibliographic metadata. We set $y(\mathcal{A}) = \texttt{True}$ for award-winning papers and sample corresponding non-award-winning papers with $y(\mathcal{A}) = \texttt{False}$. To be specific, for Best Paper Awards, the non-award-winning paper $\mathcal{A}^-$ is required to be published in the same venue as the award-winning paper $\mathcal{A}^+$. For the Nobel Prize, $\mathcal{A}^-$ is required to be authored by the same scientist as $\mathcal{A}^+$, ensuring comparability within an author's body of work. Note that the Award dimension also spans different time horizons: Best Paper Award prediction corresponds to a shorter-time horizon, as such awards are typically announced within a few months after paper acceptance. In contrast, the Nobel Prize reflects longer-term impact, given the substantial time lag between publication and the conferral of the award (Mitsis, 2022).

Patent and Media. We retrieve the records of papers referenced by patents and news/social media posts from SciSciNet (Lin et al., 2023). For other public-use dimensions, such as policy documents, the corresponding resources (Szomszor and

Adie, 2022) require restricted access and are not publicly available. Therefore, we do not include them in SciImpact. We link SciSciNet papers to MAPLE using the MAG identifier to determine the field (among the 19) to which each paper belongs. $y(\mathcal{A})$ is defined as the number of patent references and media mentions, respectively, recorded by SciSciNet.

Code. We retrieve paper-associated GitHub repositories from Papers with Code (Papers with Code, 2019), retaining those with at least 10 stars and discarding repositories with missing or extremely short README files. We collect the star count of each retrieved repository via the GitHub REST API (GitHub, 2022), which defines $y(\mathcal{A})$. The repository README serves as the primary textual input for the artifact $\mathcal{A}$.

Dataset and Model. To capture adoption in the machine learning ecosystem, we retrieve Hugging Face dataset and model cards from Yang et al. (2024b) and Liang et al. (2024), respectively. For each artifact, we collect the card text and platform-provided statistics, defining $y(\mathcal{A})$ as the corresponding download count.

3.2 Stage 2: Impact Labeling and Pair Generation

In Stage 2, after computing the impact metric $y(\mathcal{A})$ for each candidate retrieved in Stage 1, we construct contrastive pairs $(\mathcal{A}^+, \mathcal{A}^-)$ within each dimension and field. Following the dimension-specific constraints in Table 2, each pair is formed by selecting two artifacts from the same field (and matching publication year, venue, or author when applicable, as described in Section 3.1) such that $\mathcal{A}^+$ exhibits higher impact than $\mathcal{A}^-$. For count-based dimensions, both artifacts must exceed a minimum activity threshold and satisfy a minimum relative gap (e.g., $y(\mathcal{A}^+)/y(\mathcal{A}^-) \geq 2$) to ensure meaningful contrast.

3.3 Stage 3: Filtering and Quality Control

In Stage 3, we filter and sample the constructed pairs to improve text completeness, reduce noise, and balance coverage across fields. We prioritize pairs with complete textual inputs required for modeling (e.g., title and abstract for papers) and discard candidates with missing or clearly invalid text. To balance fields, we target 4,000/3,000/3,000 train/-validation/test pairs for computer science, physics, chemistry, and medicine, and 400/300/300 for each remaining field. If a field lacks enough qualified pairs, we retain all available ones. For pairs with missing text, we attempt recovery by re-fetching from online resources (e.g., Semantic Scholar; Ammar et al., 2018) using identifiers or title-based matching, keeping only reliably recovered text. Finally, we remove duplicate pairs induced by cross-source linking.

Dimension	# of Pairs	Mean Text Len.	Mean Pair Len.
Citation	43,309	156.9	313.8
Award	42,033	114.8	229.6
Patent	45,745	160.2	320.5
Media	52,739	166.8	333.6
Code	9,193	448.9	897.8
Dataset	10,517	344.8	689.5
Model	12,463	257.6	515.2

Table 3: Dataset statistics by dimension. Mean lengths are measured in word count per artifact input and per contrastive pair $(\mathcal{A}^+, \mathcal{A}^-)$, respectively.

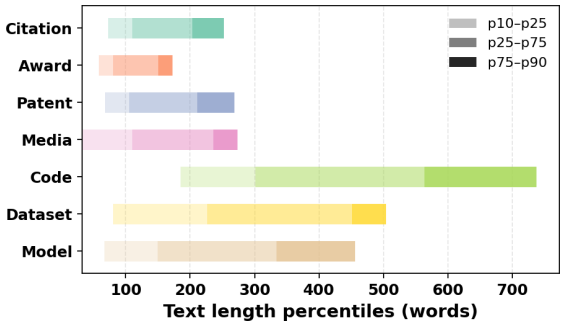

Figure 3: Text length distribution by dimension. Each horizontal bar represents percentile ranges of artifact input length (word count): p10–p25 (light), p25–p75 (medium), and p75–p90 (dark).

3.4 Dataset Statistics

Table 3 summarizes the number of contrastive pairs and the average input length (word count) for each dimension. One can observe that the mean artifact text length varies across dimensions: tasks using paper abstracts (i.e., Citation, Award, Patent, and Media) exhibit similar lengths, as abstracts are typically concise (often under 300 words). By contrast, tasks using repository or card text (i.e., Code, Dataset, and Model) have longer and more variable inputs due to the richer, heterogeneous content in README files and Hugging Face cards. Figure 3 illustrates these patterns with percentile bands of text length for each impact dimension.

4 Experiments

In this section, we comprehensively evaluate a diverse set of models, including:

	Citation	Award	Patent	Media	Code	Dataset	Model	Average
Closed-Source Models								
`GPT-4.1-mini` (Achiam et al., 2023)	0.664	0.745	0.592	0.608	0.603	0.596	0.572	0.626**
`o4-mini` (OpenAI, 2025)	0.688	0.772	0.604	0.627	**0.658**	0.589	0.557	0.642*
`Claude-haiku-4.5` (Anthropic, 2025)	0.662	0.780	0.596	0.632	0.626	0.602	0.542	0.634**
Open-Source Models								
`LLaMA-3.2-3B` (Grattafiori et al., 2024)	0.534	0.539	0.534	0.517	0.513	0.526	0.548	0.530**
`LLaMA-3-8B` (Grattafiori et al., 2024)	0.552	0.625	0.534	0.594	0.547	0.549	0.534	0.562***
`LLaMA-3.1-8B` (Grattafiori et al., 2024)	0.579	0.652	0.534	0.589	0.525	0.534	0.535	0.564***
`Qwen3-4B` (Yang et al., 2025)	0.630	0.680	0.549	0.587	0.573	0.560	0.541	0.589***
`Qwen2.5-7B` (Yang et al., 2024a)	0.601	0.646	0.557	0.604	0.563	0.592	0.560	0.589**
`Qwen2.5-14B` (Yang et al., 2024a)	0.565	0.672	0.586	0.620	0.577	0.561	0.562	0.592**
`Ministral-3-3B` (Mistral AI Team, 2025)	0.559	0.642	0.536	0.607	0.542	0.503	0.519	0.558***
`Nemotron-3-Nano-30B` (Blakeman et al., 2025)	0.537	0.618	0.500	0.504	0.528	0.565	0.549	0.543***
Fine-Tuned Models								
`SFT-LLaMA-3.2-3B`	0.653	0.806	0.629	0.697	0.618	0.618	0.625	0.664**
`SFT-Qwen3-4B`	**0.699**	**0.837**	**0.640**	**0.720**	0.626	**0.630**	**0.644**	**0.685**

Table 4: Pairwise prediction accuracy of different models across 7 impact dimensions. The **Average** column reports the average performance across all dimensions. Bold values denote the best score within each dimension. Asterisks indicate statistical significance compared to `SFT-Qwen3-4B` ($^{*}: p < 0.05$, $^{**}: p < 0.01$, $^{***}: p < 0.001$).

3 Closed-Source LLMs: `GPT-4.1-mini` (Achiam et al., 2023), `o4-mini` (OpenAI, 2025), and `Claude-haiku-4.5` (Anthropic, 2025)

8 Open-Weight LLMs: `Qwen3-4B` (Yang et al., 2025), `Qwen2.5-7B` (Yang et al., 2024a), `Qwen2.5-14B` (Yang et al., 2024a), `LLaMA-3.2-3B` (Grattafiori et al., 2024), `LLaMA-3-8B` (Grattafiori et al., 2024), `LLaMA-3.1-8B` (Grattafiori et al., 2024), `Ministral-3-3B` (Mistral AI Team, 2025), and `Nemotron-3-Nano-30B` (Blakeman et al., 2025)

2 Supervised Fine-tuned (SFT) Variants: `SFT-Qwen3-4B` and `SFT-LLaMA-3.2-3B`

4.1 Task Setup and Evaluation Protocol

We use a standardized instruction-following prompt format that (1) specifies the target impact dimension and (2) constrains the output to a strict, easily parsable form. The textual input varies by dimension: for Citation, Award, Patent, and Media, we use the paper title and abstract; for Code, Dataset, and Model, we use the corresponding repository README, Hugging Face dataset card, or Hugging Face model card, respectively. Across all dimensions, inputs are truncated to a maximum of 1,000 words when necessary to ensure consistent prompt length across instances. An example template for the Best Paper Award in computer science conferences is shown below, and full prompts for all dimensions are provided in Appendix B.

Given two scientific artifacts $(\mathcal{A}^+, \mathcal{A}^-)$ from the same field, a model is asked to predict which artifact will achieve higher future impact in a certain dimension. (Note that in all test sets, there is a 50% probability that option A in the prompt has higher impact than option B, and a 50% probability of the reverse.) We parse model outputs by exact string matching to the two allowed responses. If the output is invalid (i.e., does not exactly match either option), we retry until a valid response is obtained, ensuring that formatting issues do not affect accuracy. We report *pairwise accuracy*, defined as the percentage of instances in which the model correctly identifies $\mathcal{A}^+$ over $\mathcal{A}^-$.

System: You are an impartial judge deciding which of two papers won the Best Paper Award. Your reply must be exactly one sentence and must be one of these two options:
– Paper A won the Best Paper Award
– Paper B won the Best Paper Award
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper won the Best Paper Award?
Reply with exactly one sentence following the system instruction.

4.2 Supervised Fine-tuning Setup

To investigate whether task-specific training improves models' performance in scientific impact prediction, we perform SFT on two representative open-weight LLMs: `Llama-3.2-3B` (Grattafiori et al., 2024) and `Qwen3-4B` (Yang et al., 2025). Both models are fine-tuned on the training split aggregated across all impact dimensions and fields, and hyperparameters are selected based on performance on the corresponding aggregated validation split. Full-parameter fine-tuning is conducted us-

	Comp. Sci.	Physics	Chemistry	Medicine	Other Fields	Average
Closed-Source Models						
GPT-4.1-mini (Achiam et al., 2023)	0.625	0.706	0.685	0.673	0.586	0.655*
o4-mini (OpenAI, 2025)	0.639	**0.730**	0.690	0.710	0.617	0.677
Claude-haiku-4.5 (Anthropic, 2025)	0.631	0.680	0.694	0.730	0.615	0.670*
Open-Source Models						
LLaMA-3.2-3B (Grattafiori et al., 2024)	0.533	0.541	0.528	0.525	0.531	0.532***
LLaMA-3-8B (Grattafiori et al., 2024)	0.561	0.616	0.585	0.556	0.551	0.574**
LLaMA-3.1-8B (Grattafiori et al., 2024)	0.560	0.644	0.607	0.552	0.559	0.584**
Qwen3-4B (Yang et al., 2025)	0.590	0.653	0.633	0.607	0.577	0.612**
Qwen2.5-7B (Yang et al., 2024a)	0.587	0.661	0.617	0.579	0.565	0.602**
Qwen2.5-14B (Yang et al., 2024a)	0.597	0.684	0.594	0.597	0.585	0.612*
Ministral-3-3B (Mistral AI Team, 2025)	0.547	0.619	0.616	0.577	0.556	0.583***
Nemotron-3-Nano-30B (Blakeman et al., 2025)	0.544	0.525	0.578	0.533	0.510	0.538***
Fine-Tuned Models						
SFT-LLaMA-3.2-3B	0.652	0.700	0.718	0.722	0.681	0.695*
SFT-Qwen3-4B	**0.669**	0.717	**0.768**	**0.743**	**0.704**	**0.720**

Table 5: Pairwise prediction accuracy of different models across scientific fields. The **Average** column reports the average performance across all fields. Bold values denote the best score within each field. Asterisks indicate statistical significance compared to SFT-Qwen3-4B ($^{*}: p < 0.05$, $^{**}: p < 0.01$, $^{***}: p < 0.001$).

ing LLaMA-Factory[1]. All SFT experiments are performed on four NVIDIA H20 GPUs. Complete training commands and additional implementation details are provided in Appendix C.

4.3 Main Results

Tables 4 and 5 present the detailed performance of all models. We also computed the averages across all dimensions and fields, and performed pairwise t-tests between each model and SFT-Qwen3-4B. Statistical significance is indicated in both tables. Based on these results, we highlight three key observations.

Effectiveness of Training on SciImpact. Fine-tuned models substantially outperform their corresponding base models, demonstrating that SciImpact provides a strong supervision signal for learning impact-relevant cues from artifact text. Notably, SFT-Qwen3-4B achieves the best performance on nearly all impact dimensions and across almost every academic field; the only exceptions are the Code dimension and the Physics field, where o4-mini attains the highest scores. Similarly, SFT-LLaMA-3.2-3B consistently outperforms all non-fine-tuned open-weight models and surpasses several closed-source systems in specific settings, including the Patent and Media dimensions and the Computer Science field. Overall, these results indicate that relatively small open models, when fine-tuned on SciImpact, can outperform widely used instruction-tuned baselines and compete effectively with larger open-source LLMs and much stronger closed-source models.

Importance of Dimension- and Field-Specific Evaluation. Statistical analysis using a two-factor ANOVA without replication on Tables 4 and 5 reveals significant performance variation across both impact dimensions and scientific fields.

Among all dimensions, Award yields the highest prediction accuracy and is significantly easier than every other task ($p < 0.001$). This trend aligns with the nature of the underlying signal: unlike many long-term impact measures that accumulate gradually, Best Paper Awards (in computer science conferences and MDPI journals) are typically determined by expert committees within a relatively short time frame, often months to one year. As a result, such awards emphasize salient, human-recognizable indicators of excellence that are more readily detectable from artifact text. Notably, Nobel Prize related signals constitute an exception with respect to time scale. They usually reflect research that has been validated over decades, yet they achieve even higher accuracy across models as shown in Table 6. One plausible explanation is that Nobel associated work in the natural sciences is more likely to contain explicit textual markers of discovery, such as named compounds, experimental protocols, or clinical trial outcomes. In contrast, impact in computer science may depend more heavily on rapidly evolving trends and community dynamics that are difficult to infer from text alone. In addition, Nobel Prize related papers are often widely discussed and extensively cited, making them more likely to appear in LLM

[1] https://github.com/hiyouga/LLaMA-Factory

Model Group	Best Paper Award	Nobel Prize
Base Models	0.606	0.737
SFT Models	0.748	0.898

Table 6: Average pairwise accuracy across Award subtypes. Best Paper Award aggregates the Award dimension from Computer Science and Other fields (i.e., MDPI journals), while Nobel Prize spans Physics, Chemistry, and Medicine.

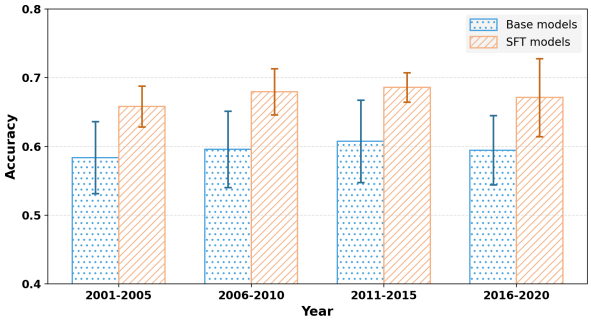

Figure 4: Citation accuracy by publication year. The bar chart compares the average accuracy of Base models (blue, dotted hatch) and SFT models (orange, diagonal hatch) across four five-year intervals from 2001 to 2020. The error bars represent the standard deviation of accuracy across models within each time bin.

pre-training corpora. This broader exposure may confer partial prior knowledge of canonical Nobel winning contributions and further reduce the difficulty of identifying higher impact artifacts in paired comparisons.

Motivated by this field effect within the Award dimension, we further analyze field-wise differences across all impact dimensions and find that Computer Science and the aggregated Other Fields are significantly more challenging than Chemistry, Medicine, and Physics, with all 2×3 pairwise comparisons yielding $p < 0.01$.

It is precisely these systematic variations in prediction difficulty across impact dimensions and scientific fields that motivate our construction of a multi-dimensional, multi-field benchmark for scientific impact prediction.

Challenges in Predicting Long-Term Impact. For most impact dimensions other than awards, model performance remains modest and only weakly differentiated, with accuracies largely clustered between 0.50 and 0.65. This plateau highlights the intrinsic difficulty of forecasting long-term scientific impact from textual content alone, even for strong contemporary LLMs.

This challenge is further illustrated by the citation analysis in Figure 4, which reports citation prediction accuracy across four five-year publication intervals from 2001 to 2020. Performance remains largely stable over time for both base models and SFT models, with no clear advantage for older papers whose long-term citation trajectories have fully materialized. This temporal invariance is primarily a consequence of our input design: models observe only titles and abstracts, without access to temporal cues such as publication year, citation histories, or early reception signals. Consequently, age-related information is removed, forcing models to rely solely on intrinsic textual signals and resulting in comparable difficulty across publication periods.

Although SFT models consistently outperform their base counterparts in every time bin, the persistent performance plateau suggests that many high-impact papers do not exhibit easily separable signals at the time of publication. Taken together, these findings underscore the fundamental uncertainty of long-term impact prediction from early textual descriptions and motivate the need for benchmarks such as SciImpact that explicitly stress-test content-based reasoning while providing stronger supervision for learning subtle, impact-relevant cues.

5 Conclusion

We introduce SciImpact, a multi-dimensional, multi-field benchmark for scientific impact prediction, spanning the Citation, Award, Patent, Media, Code, Dataset, and Model dimensions, as well as fields including Computer Science, Physics, Chemistry, Medicine, and various other areas. Our evaluation of 11 LLMs demonstrates the heterogeneous nature of scientific impact: models perform best on the Award dimension, where textual cues closely align with evaluative criteria, while dimensions such as Patent and Media remain challenging due to latent external factors (e.g., market timing and societal relevance). Task-specific training on SciImpact proves highly effective, with relatively small models like `SFT-Qwen3-4B` consistently outperforming larger open-source baselines and even stronger closed-source models. Observed performance patterns across dimensions and fields reveal where textual signals are sufficient and where additional context may be necessary. Overall, SciImpact provides a rigorous platform for evaluating and improving multi-dimensional, multi-field scientific impact prediction, supporting the development of more effective and generalizable models.

Limitations

Despite the usefulness of SCIIMPACT as a benchmark and the demonstrated efficacy of LLMs fine-tuned on it, there remain the following two limitations that motivate future work.

Text Truncation and Scope. We limit textual input to at most 1,000 words for all dimensions to maintain consistent prompt lengths across instances. Consequently, SCIIMPACT provides models with a truncated view of papers and long-form artifacts such as extended READMEs or dataset / model cards. While additional context could offer more insights into an artifact's novelty, technical depth, and potential for downstream adoption, incorporating full-length content is constrained by current computational budgets and the practical challenges of long-context inference for LLMs, including higher latency, increased cost, and potential degradation when reasoning over very long inputs. Future work could explore long-context models or hierarchical and retrieval-based reading strategies to make fuller use of artifact content while maintaining scalability.

Pairwise Simplification. We formulate impact prediction as a pairwise contrastive task, in which models determine whether $\mathcal{A}^+$ is more impactful than $\mathcal{A}^-$. This design provides a clean and robust evaluation of discriminative ability and mitigates scale differences across dimensions. At the same time, it abstracts away some of the complexities of real-world impact forecasting. In practice, users may also be interested in absolute estimates, such as predicting citation counts, or global rankings over large candidate pools. Accordingly, while strong performance on SCIIMPACT demonstrates useful discriminative capabilities, further work is needed to extend these methods to calibrated regression predictions or large-scale ranking scenarios.

Ethical Considerations

A central ethical consideration of this work concerns the risk posed by Goodhart's Law (Strathern, 1997). If predictive models of scientific impact are deployed in high-stakes settings such as funding allocation, hiring, or promotion, researchers may be incentivized to optimize their writing style, topical choices, or dissemination strategies to satisfy the model's signals rather than to improve the intrinsic quality of their scientific contributions. Such feedback loops could distort research behavior and narrow the diversity of scientific inquiry. We therefore emphasize that SCIIMPACT and models trained on it are intended strictly as decision-support and filtering tools to assist human discovery and exploration, not as autonomous or authoritative systems for scientific evaluation.

Relatedly, models trained to predict multi-dimensional impact may inherit biases present in historical data and public signals. For example, award recognition, media attention, or artifact adoption may systematically favor certain fields, institutions, or research communities, potentially reinforcing existing inequities in science. While SCIIMPACT broadens impact beyond citations and covers a wide range of fields, it does not eliminate such structural biases. Users of this benchmark and resulting models should therefore exercise caution, conduct bias analyses across dimensions and fields, and avoid interpreting predictions as normative judgments of scientific merit.

Moreover, our work relies on LLMs, which may further amplify societal or cultural biases encoded in their pre-training data (Bender et al., 2021). Although our goal is not to improve LLM generation, but to evaluate and fine-tune models for impact prediction, the resulting systems may still produce inconsistent or biased reasoning. In practical use, predictions should be complemented with existing bias mitigation, auditing, and calibration techniques, and should always be contextualized by domain experts.

Overall, we view scientific impact prediction as an inherently subjective and multi-faceted task. Our benchmark is designed to facilitate research into this complexity, not to replace expert judgment. Responsible use of SCIIMPACT requires maintaining human oversight, transparency about limitations, and restraint in applying model predictions to consequential decisions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, and 1 others. 2018. Construction of the literature graph in semantic scholar. In *NAACL'18*, pages 84–91.

Anthropic. 2025. System card: Claude haiku 4.5.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT'21*, pages 610–623.

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, and Aleksander Ficek. 2025. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*.

Carlos Castillo, Debora Donato, and Aristides Gionis. 2007. Estimating number of citations using author reputation. In *International Symposium on String Processing and Information Retrieval*, pages 107–117.

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *JCDL'14*, pages 351–360.

Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2015. Will this paper increase your h-index? scientific impact prediction. In *WSDM'15*, pages 149–158.

Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A century of science: Globalization of scientific collaborations, citations, and innovations. In *KDD'17*, pages 1437–1446.

Lawrence D Fu and Constantin Aliferis. 2008. Models for predicting and explaining citation count of biomedical articles. In *AMIA'08*, page 222.

GitHub. 2022. Github rest api documentation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xuemei Gu and Mario Krenn. 2024. Impact4cast: forecasting high-impact research topics via machine learning on evolving knowledge graphs. In *ICML 2024 AI for Science Workshop*.

Jun Hirako, Ryohei Sasano, and Koichi Takeda. 2023. Realistic citation count prediction task for newly published papers. In *Findings of EACL'23*, pages 1131–1141.

Jeff Huang. 2023. Best paper awards in computer science (since 1996).

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

Ching Jin, Yifang Ma, and Brian Uzzi. 2021. Scientific prizes and the extraordinary growth of scientific topics. *Nature Communications*, 12(1):5619.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *NeurIPS'21*.

Simon Koch, David Klein, and Martin Johns. 2024. The fault in our stars: An analysis of github stars as an importance metric for web source code. In *Workshop on Measurements, Attacks, and Defenses for the Web*.

Jichao Li, Yian Yin, Santo Fortunato, and Dashun Wang. 2019a. A dataset of publication records for nobel laureates. *Scientific Data*, 6(1):33.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. 2019b. A neural citation count prediction model based on peer review text. In *EMNLP'19*, pages 4914–4924.

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic analysis of 32,111 ai model cards characterizes documentation practice in ai. *Nature Machine Intelligence*, 6(7):744–753.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Mingfei Lu, Mengjia Wu, Jiawei Xu, Weikai Li, Feng Liu, Ying Ding, Yizhou Sun, Jie Lu, and Yi Zhang. 2025. From newborn to impact: Bias-aware citation prediction. *arXiv preprint arXiv:2510.19246*.

MDPI. 2025. Mdpi awards.

Mistral AI Team. 2025. Ministral 3: Strong edge-ready ai.

Pandelis Mitsis. 2022. The nobel prize time gap. *Humanities and Social Sciences Communications*, 9(1):407.

OpenAI. 2025. Openai o3 and o4-mini system card.

Papers with Code. 2019. Links between papers and code.

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

Filippo Radicchi, Alexander Weissman, and Johan Bollen. 2017. Quantifying perceived impact of scientific publications. *Journal of Informetrics*, 11(3):704–712.

Leiming Ren, Shimin Shan, Xiujuan Xu, and Yu Liu. 2020. Starin: An approach to predict the popularity of github repository. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 258–273.

Kiyan Rezaee, Morteza Ziabakhsh, Niloofar Nikfarjam, Mohammad M Ghassemi, Yazdan Rezaee Jouryabi, Sadegh Eskandari, and Reza Lashgari. 2025. Fos: A large-scale temporal graph benchmark for scientific interdisciplinary link prediction. *arXiv preprint arXiv:2511.18631*.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *ACL'18*, pages 87–92.

Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239.

Marilyn Strathern. 1997. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321.

Martin Szomszor and Euan Adie. 2022. Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, 3(3):624–650.

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

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

Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. 2016. On modeling and predicting individual paper citation count over time. In *IJCAI'16*, pages 2676–2682.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xinyu Yang, Weixin Liang, and James Zou. 2024b. Navigating dataset documentations in ai: A large-scale analysis of dataset cards on huggingface. In *ICLR'24*.

Yian Yin, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F Jones. 2022. Public use and public funding of science. *Nature Human Behaviour*, 6(10):1344–1350.

Sha Yuan, Jie Tang, Yu Zhang, Yifan Wang, and Tong Xiao. 2018. Modeling and predicting citation count via recurrent neural network with long short-term memory. *arXiv preprint arXiv:1811.02129*.

Yu Zhang. 2025. Internal and external impacts of natural language processing papers. In *ACL'25*, pages 488–494.

Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, and Jiawei Han. 2023. The effect of metadata on scientific literature tagging: A cross-field cross-model study. In *WWW'23*, pages 1626–1637.

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 *AAAI'25*, pages 1183–1191.

A Comparison across Fields

Figure 5 provides a field-wise view of the same three representative models examined in Figure 1, illustrating how performance varies across scientific fields and how SFT affects cross-field generalization.

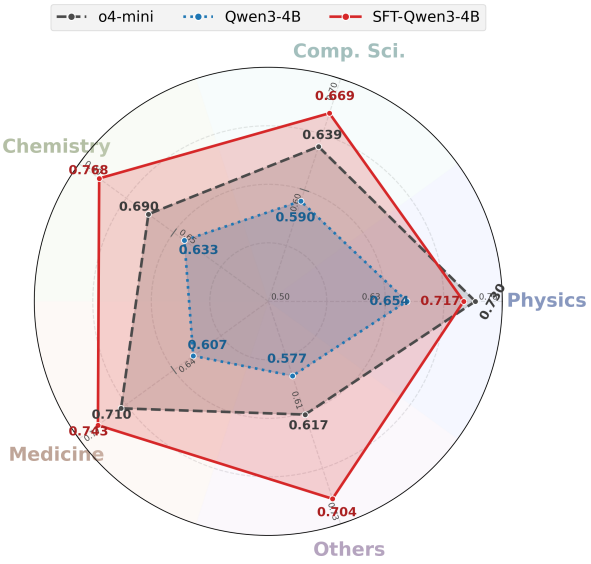

Figure 5: Performance of `o4-mini`, off-the-shelf `Qwen3-4B`, and supervised fine-tuned `Qwen3-4B` across scientific fields on SciImpact. SFT substantially enhances a 4B open-weight model's ability to predict scientific impact across all fields, enabling it to rival or surpass stronger closed-source models.

B Prompts

B.1 Citation

System: You are an impartial judge deciding which of two research papers has more citations. Your reply must be exactly one sentence and must be one of these two options:
– Paper A has more citations
– Paper B has more citations
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based solely on the information above, which paper do you think has more citations?
Reply with exactly one sentence following the system instruction.

B.2 Best Paper Award (CS Conferences)

System: You are an impartial paper reviewer. Given the titles and abstracts of two papers, identify which paper won the Best Paper award. Your reply must be exactly one sentence and must be one of these two options:
– Paper A won the best paper award.
– Paper B won the best paper award.
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper should win the Best Paper award?
Reply with exactly one sentence following the system instruction.

B.3 Best Paper Award (MDPI Journals)

System: You are an impartial judge deciding which of two MDPI papers won the MDPI Best Paper Award. Your reply must be exactly one sentence and must be one of these two options:
– Paper A won the MDPI Best Paper Award
– Paper B won the MDPI Best Paper Award
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper won the MDPI Best Paper Award?
Reply with exactly one sentence following the system instruction.

B.4 Nobel Prize

System: You are an impartial judge deciding which of two research papers is the Nobel prize-winning paper. Your reply must be exactly one sentence and must be one of these two options:
– Paper A is the Nobel prize-winning paper.
– Paper B is the Nobel prize-winning paper.
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper is the Nobel prize-winning paper?
Reply with exactly one sentence following the system instruction.

B.5 Patent

System: You are an impartial judge deciding which of two research papers would be cited in more patents. Your reply must be exactly one sentence and must be one of these two options:
– Paper A could be cited in more patents.
– Paper B could be cited in more patents.
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper could be cited in more patents?
Reply with exactly one sentence following the system instruction.

B.6 Media

System: You are an impartial judge deciding which of two research papers would be cited in more media mentions. Your reply must be exactly one sentence and must be one of these two options:
– Paper A could get more media mentions.
– Paper B could get more media mentions.
You are not allowed to output anything else—no explanations, no extra words.

User: Paper A: `<artifact text for A>`
Paper B: `<artifact text for B>`

Based on the information above, which paper could get more media mentions?
Reply with exactly one sentence following the system instruction.

B.7 Code

System: You are an impartial judge deciding which of two GitHub repositories has more stars. Your reply must be exactly one sentence and must be one of these two options:
– GitHub repo A has more stars.
– GitHub repo B has more stars.
You are not allowed to output anything else—no explanations, no extra words.

User: GitHub repo A README: `<artifact text for A>`
GitHub repo B README: `<artifact text for B>`

Based on the information above, which repository has more stars?
Reply with exactly one sentence following the system instruction.

B.8 Dataset

System: You are an impartial judge deciding which of two Hugging Face datasets has more downloads. Your reply must be exactly one sentence and must be one of these two options:
– Dataset A has more downloads.
– Dataset B has more downloads.
You are not allowed to output anything else—no explanations, no extra words.

User: Dataset A: `<artifact text for A>`
Dataset B: `<artifact text for B>`

Based on the information above, which dataset has more downloads?
Reply with exactly one sentence following the system instruction.

B.9 Model

System: You are an impartial judge deciding which of two Hugging Face models has more downloads. Your reply MUST be exactly one sentence and must be one of these two options:
– Model A has more downloads.
– Model B has more downloads.
You are not allowed to output anything else—no explanations, no extra words.

User: Model A: `<artifact text for A>`
Model B: `<artifact text for B>`

Based on the information above, which model has more downloads?
Reply with exactly one sentence following the system instruction.

C Implementation Details for Supervised Fine-tuning

We conduct SFT using the LLaMA-Factory framework.[2] Both `Qwen3-4B` and `LLaMA-3.2-3B` are fine-tuned under an identical training configuration, differing only in the base model checkpoint and the prompt template.

Models are trained on the aggregated training split spanning all 19 scientific fields and seven impact dimensions, and validated on the corresponding aggregated validation split. Each training instance consists of a single instruction-following prompt formatted as described in Appendix B, with a binary forced-choice output. We conduct full-parameter fine-tuning using the following hyperparameters:

- Learning rate: `2e-5`
- Epochs: `1` (`Qwen3-4B`), `3` (`LLaMA-3.2-3B`)
- Per-device batch size: `8` (train / eval)
- Gradient accumulation steps: `2`
- Effective batch size: `64`
- Learning rate schedule: cosine
- Warmup ratio: `0.1`

We set the maximum input length to 4,096 tokens, truncating longer inputs accordingly. Mixed-precision training with bf16 is enabled, and FlashAttention with SDPA is used to improve memory efficiency. All experiments are conducted on four NVIDIA H20 GPUs, leveraging DeepSpeed ZeRO Stage 2 for memory optimization. To mitigate memory fragmentation during long-context training, we enable expandable CUDA memory segments. Evaluation is performed on the validation split every 500 training steps.

[2] https://github.com/hiyouga/LLaMA-Factory