
ForeCite: Adapting Pre-Trained Language Models to Predict Future Citation Rates of Academic Papers

Gavin Hull

Department of Computer Science, Department of Mathematics and Statistics
Memorial University of Newfoundland and Labrador
St. John's, NL A1C 5S7
ghull@mun.ca

Alex Bihlo

Department of Mathematics and Statistics
Memorial University of Newfoundland and Labrador
St. John's, NL A1C 5S7
abihlo@mun.ca

Abstract

Predicting the future citation rates of academic papers is an important step toward the automation of research evaluation and the acceleration of scientific progress. We present **ForeCite**, a simple but powerful framework to append pre-trained causal language models with a linear head for average monthly citation rate prediction. Adapting transformers for regression tasks, ForeCite achieves a test correlation of $\rho = 0.826$ on a curated dataset of 900K+ biomedical papers published between 2000 and 2024, a 27-point improvement over the previous state-of-the-art. Comprehensive scaling-law analysis reveals consistent gains across model sizes and data volumes, while temporal holdout experiments confirm practical robustness. Gradient-based saliency heatmaps suggest a potentially undue reliance on titles and abstract texts. These results establish a new state-of-the-art in forecasting the long-term influence of academic research and lay the groundwork for the automated, high-fidelity evaluation of scientific contributions.¹

1 Introduction

This paper explores the potential of pre-trained language models to predict future citation rates directly from the textual content of scientific manuscripts. Though citation rates serve only as a weak proxy for true scholarly impact, they offer a tractable signal through which to examine how contemporary models might begin to interpret and anticipate patterns of scientific attention at scale. By framing citation prediction as a text-to-signal learning problem, this work establishes a foundation for more context-aware, adaptive systems that could assist in analyzing—and eventually guiding—the evolving landscape of scientific research.

Traditional citation-prediction methods depend heavily on external metadata—author profiles, citation graphs, or peer-review excerpts—to anticipate future citations [Pobiedina and Ichise, 2016, Davletov et al., 2014, Li et al., 2019]. While these features can yield reasonable forecasts in narrow domains, they inherently overlook the semantic information and methodological quality embedded in the manuscript itself. In contrast, modern language models have been shown to capture the deep contextual and conceptual relationships in text [Brown et al., 2020], although their application to

¹Relevant code will be made available upon publication.

citation forecasting has largely been limited to small text fragments or as embeddings for separate downstream learners [Hirako et al., 2024, 2023, van Dongen et al., 2020, Wenniger et al., 2023].

To bridge this gap, we introduce ForeCite, a plug-and-play framework for adapting pre-trained causal transformers for regression tasks, enabling end-to-end prediction of average monthly citation rates. We evaluate ForeCite on a curated corpus of 900K+ medical articles published between 2000 and 2024, demonstrating superior predictive performance over state-of-the-art baselines. Our extensive suite of experiments (including scaling-law analysis across model sizes and data volumes, temporal holdout studies to probe concept drift, gradient-based saliency visualizations for interpretability, and targeted ablations) evidence the potential of large language models (LLMs) in scientific impact forecasting.

We make four main contributions:

1. **Framework for regression.** A simple general method for adapting pre-trained causal language models for regression tasks while preserving semantic knowledge.
2. **New state-of-the-art.** A Spearman rank correlation of $\rho = 0.826$ between predicted and true citation rates, marking a 27-point improvement over the previous state-of-the-art.
3. **Comprehensive testing.** In-depth analysis of scaling behavior, temporal robustness, interpretability, and targeted ablations.
4. **Open science commitment.** Public release of code and model checkpoints of various sizes to spur further advancements in this field.

2 Related work

Citation prediction. Early work framed citation forecasting as a classification or simple regression problem, using handcrafted features derived from author and venue metadata [Ibáñez et al., 2009, Chakraborty et al., 2014, Pobiedina and Ichise, 2016]. Alternative graph-based methods exploited network structure, both temporal and topological, to extend forecasting horizons [Davletov et al., 2014]. Some efforts also employed manual mathematical modeling to capture citation dynamics more explicitly [Wang et al., 2013].

More recent neural models incorporate textual and early-citation signals via multilayer perceptrons or recurrent neural networks, yielding improved but still limited performance [Ruan et al., 2020, Jamal et al., 2024]. The latest transformer-based predictors leverage pre-trained language embeddings over abstracts or document chunks, but typically only use them as fixed features for downstream regressors rather than in an end-to-end predictive framework [Hirako et al., 2024, van Dongen et al., 2020, Hirako et al., 2023, Wenniger et al., 2023]. A more in-depth analysis of techniques used in the field can be found in Aiza et al. [2024].

Large language models. The transformer architecture, built solely on self-attention and forgoing recurrence, has become a foundational tool in natural language processing (NLP), demonstrating superior performance and parallelizability across tasks [Vaswani et al., 2023]. Scaling transformers to billions of parameters has yielded emergent capabilities, including few-shot learning, that have redefined model generality and adaptability [Brown et al., 2020]. Despite these advances, fine-tuning causal transformers for regression tasks such as citation rate prediction remains largely underexplored in the literature.

Scaling-laws. Empirical studies reveal that language models’ performance follows predictable power-law relationships with respect to both model size and training data volume [Kaplan et al., 2020]. While these scaling-laws extend to many downstream tasks, the precise improvement achievable through fine-tuning is task dependent. Our work is the first to systematically characterize these trends for citation rate prediction.

Interpretability. Understanding which textual elements drive model predictions is critical for transparency and interpretability. Interpretability techniques for transformer-based architectures fall into several categories. Attention-based visualization remains a popular means of attributing model decisions, highlighting tokens which receive high self-attention scores [Hao et al., 2021]. Gradient-based saliency methods, originally developed for vision models and later adapted for NLP, assign

importance scores to input tokens via back-propagated gradients [Li et al., 2016]. We apply these techniques here to highlight the impact of individual tokens on predicted citation rates.

3 Methods

3.1 Data collection & preprocessing

We constructed our dataset by querying the Elsevier API for seven biomedical keywords across every month from January 2000 to December 2024 with a maximum of 6K resulting articles per query. After de-duplication, our dataset underwent rigorous filtering of corrupted entries, non-English and restricted-access items, and outliers via several criteria (e.g. article length, completeness, etc.). This process yielded a high-quality corpus comprising 900K+ articles (see Figure 1), each annotated with its publication date and total number of citations as of December 2024.

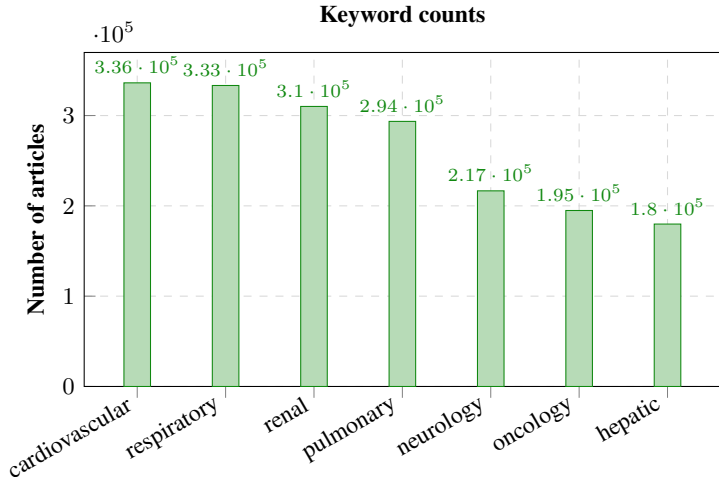


Figure 1: Distribution of articles in the final dataset containing each domain-specific keyword.

Aligning with standard practice in citation prediction [Hirako et al., 2024, van Dongen et al., 2020, Hirako et al., 2023, Wenniger et al., 2023], we applied a log-transform to the average monthly citation rates. This produced a near-Gaussian distribution (Figure 2), motivating the standardization of our data prior to training. The mean and standard deviation were calculated using the training split to prevent leakage. Given the large quantities of parameters aggregated during LLM inference, the central limit theorem suggests that this transformation should encourage more stable optimization.

Each article, originally in XML format, was converted to a consistent Markdown representation via a custom XSLT script. This standardized format preserved structural elements such as section headings, section content, and figure/table captions, as well as available metadata such as journal name and copyright, enabling consistent tokenization across all models.

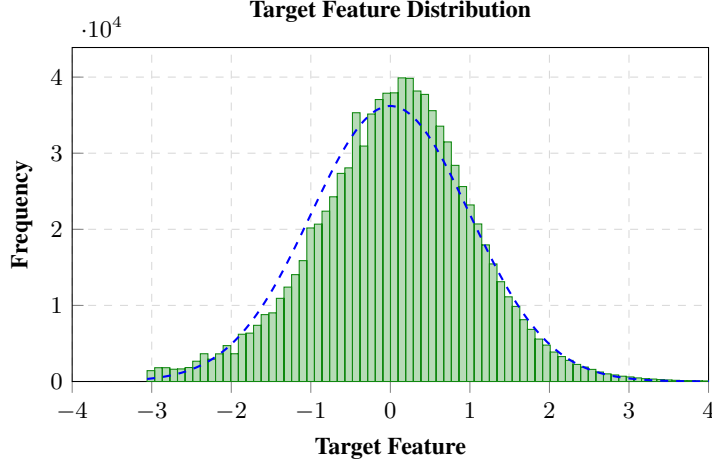


Figure 2: Standardized average monthly citation rates post log-transform, overlaid with a scaled standard normal curve for reference.

3.2 Model architecture

All experiments followed an identical architectural pipeline: we appended a simple linear head to the final hidden state of a pre-trained Hugging Face `AutoModelForCausalLM` model serving as the base language model. All other architectural decisions (e.g. tokenization scheme, context window length, etc.) were inherited from the base model configurations. To maximize hardware utilization, we apply 4-bit NF4 quantization and `bfloat16` arithmetic; this configuration enabled highly efficient training, allowing even our largest models to fit within our memory budget without sacrificing predictive fidelity.

3.3 Training procedure

Each model was trained in two phases:

1. **Initial training phase:** The base LLM was frozen while only the appended linear head was trained.
2. **Fine-tuning phase:** The base LLM was unfrozen and jointly fine-tuned with the linear head via QLoRA [Dettmers et al., 2023].

All training runs used a 90/10 train/test split. A complete list of training and fine-tuning hyperparameters is provided in Table 6 in Appendix A. Any parameter not listed was left at its default value.

All experiments were conducted on Compute Canada’s *Narval* cluster, using $6 \times$ AMD EPYC 7413 CPUs and $2 \times$ NVIDIA A100 SXM4 GPUs. Although each run’s wall-clock time varied by model size and data volume, we exclude these details here for brevity. All experiments had a maximum run-time of 7 days.

3.4 Experimental setup

In the section that follows, we present the results of our two primary experiments:

Scaling-law analysis. To characterize how the performance of our framework scales with model size and data volume, we trained five transformer variants (0.5B – 14B parameters) across five fractions of our corpus (1% – 16%). Data volumes never exceeded 16% of our corpus due to computational constraints. A full list of the models used in this work, alongside their publishers and parameter counts, can be found in Appendix B, Table 7. All evaluations were—unless otherwise specified—performed on a randomly generated hold-out test set. A full list of our evaluation metrics can be found in Appendix C.

Gradient saliency analysis. To reveal which textual features drive ForeCite’s predictions, we computed gradient-based attributions over a number of samples [Li et al., 2016]. For each sample, we accumulated token-level gradients of the predicted score with respect to the input embeddings, quantifying the impact of individual tokens on predicted citations.

Temporal holdout. To verify the robustness of our framework in real-world applications, we trained Bloom-560m on a random subset of articles published before 2023 (equivalent to 16% of our total corpus). Evaluating on monthly cohorts thereafter, we reported a rolling Pearson correlation. The Bloom-560m model was selected due to its release date of 2022, ensuring that no test data could have been seen during its pre-training.

4 Results

4.1 Scaling-law analysis

Figure 3 presents a 3D surface of Pearson correlation (r) over model sizes and data volumes. Our largest configuration, Qwen2.5-14B on 16% data, attained $r = 0.844$, whereas our smallest, Bloom-560m on 1% data, reported nearly half: $r = 0.430$. Figures 4a and 4b isolate these trends:

1. As data volumes increase, all models improved monotonically, with steeper gains being recorded for the smaller models.
2. As model size increases, performance generally rose, though mid-sized models (1B – 8B) showed noticeable deviations.

A complete grid of all 25 values for each of our evaluation metrics is provided in Appendix D, Table 8.

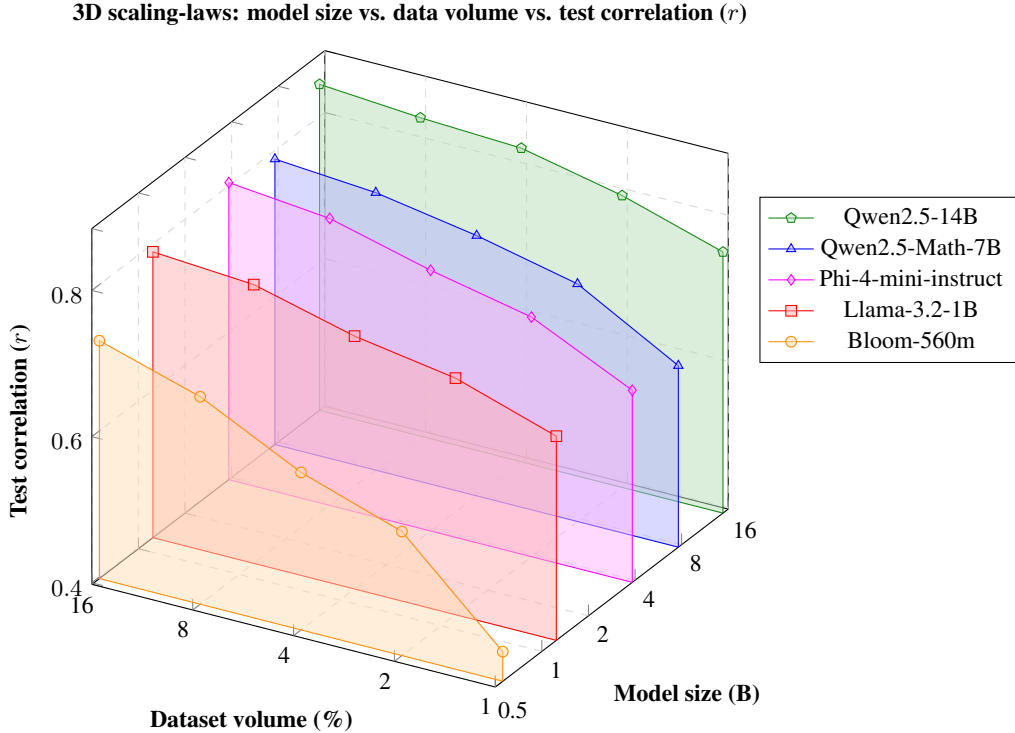
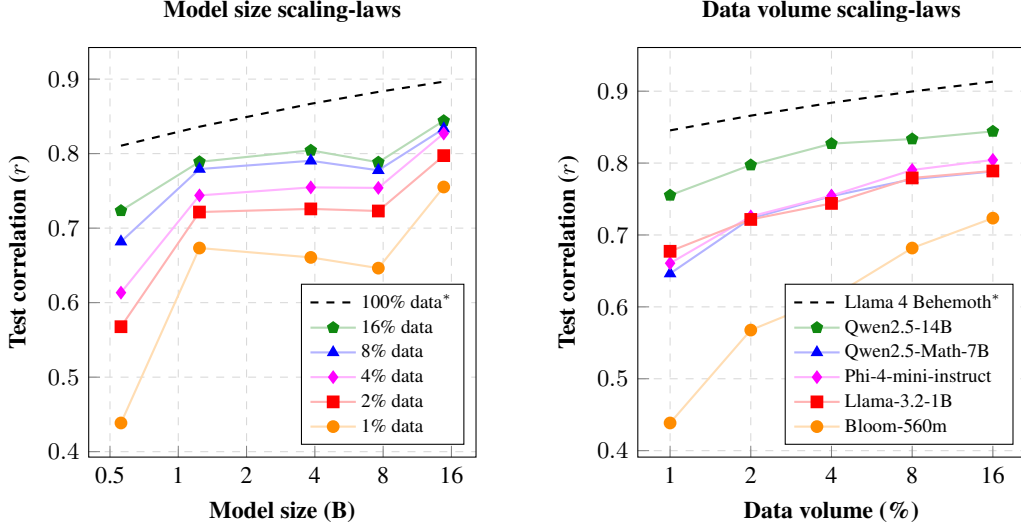


Figure 3: A 3D visualization of the relationship between model size (in billions of parameters, \log_2 scaled), data volume (in percent of total corpus, \log_2 scaled), and performance (in Pearson correlation, evaluated on a holdout test set.)



(a) A 2D visualization of the relationship between model size (in billions of parameters, \log_2 scaled) and performance (in Pearson correlation, evaluated on a holdout test set) at various data volumes (in percent of total corpus.)

(b) A 2D visualization of the relationship between data volume (in percent of total corpus, \log_2 scaled) and performance (in Pearson correlation, evaluated on a holdout test set) at various model sizes (in billions of parameters.)

*Indicates a theoretical estimate computed using our least squares model.

To anticipate the trajectory of state-of-the-art architectures trained on our full corpus, we fit a bounded least squares model of the form

$$f(p, d) = \tanh(\beta_0 + \beta_1 \log_2(p) + \beta_2 \log_2(d)), \quad (1)$$

where p is model size in billions of parameters, and d is the percent of total data used. Extrapolating to the latest state-of-the-art Llama model, Llama 4 Behemoth, with 288 B active parameters and trained on our full corpus, we estimated a Pearson correlation of $r(288, 100) = 0.9413$ and Spearman rank correlation of $\rho(288, 100) = 0.9325$. A full collection of the derived β -parameters can be found in Appendix E, Table 9. Both models exhibited an MAE of approximately 0.03.

4.2 Gradient saliency analysis

On average, title and abstract tokens exhibited the highest attribution, with general body text receiving minimal weight. Across the sampled papers, models consistently favored focused and information-dense content that reflected the main topic of each paper. For example, "empty," meta-discursive statements such as "In this article we..." were routinely penalized, whereas discourse markers indicating new information such as "additionally" and "moreover," were often rewarded (albeit not as strongly as "empty" statements were penalized). Model attribution was inconsistent across most metadata—including publication dates and journal names—only consistently using author names for the most prominent authors. Figure 10 in Appendix F displays a representative heatmap over the abstract of a sample paper [Arsenault et al., 2022].

4.3 Temporal holdout

Figure 5 charts model performance out-of-distribution from January 2023 through April 2025. The initial correlation of $r = 0.713$ closely matched the static test performance of $r = 0.721$ reported during our scaling-law analysis. Performance dropped to $r = 0.631$ after the first month, and then steadily decayed to $r = 0.511$ by mid-2025.

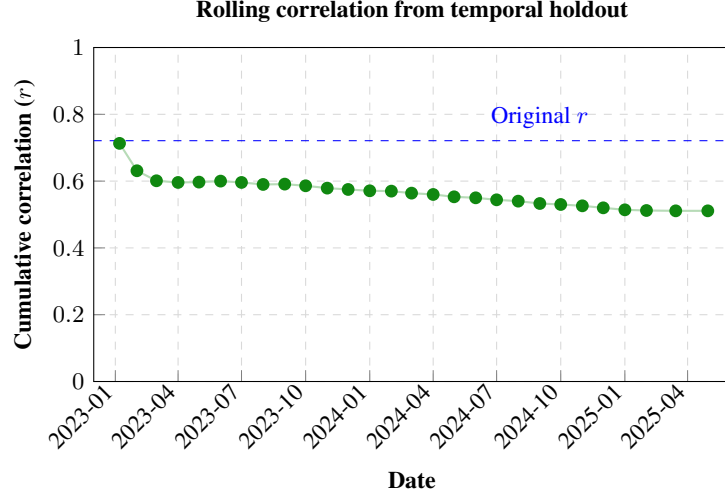


Figure 5: This graph visualizes the rolling test correlation of the Bloom-560m model, evaluated with temporal holdout. For example, the point with x value 2024-10 and y value 0.530 represents a 0.530 Pearson correlation between the true and predicted target values between 2023-01 and 2024-10. The blue line denotes the original r found during the scaling-law evaluation of Bloom-560m.

5 Discussion

5.1 Interpretation of experimental findings

Comparison to prior work. Our best models represent a substantial advancement over previous state-of-the-art techniques. SCHuBERT and MultiSCHuBERT report coefficients of determination $R^2 = 0.4$ and $R^2 = 0.454$ respectively, despite the latter using the additional visual information embedded in the figures and layout of the articles [van Dongen et al., 2020, Wenniger et al., 2023]. In comparison, we record a coefficient of determination $R^2 = 0.706$, given only textual information. Current state-of-the-art models such as CiMaTe achieve a Spearman rank correlation of $\rho = 0.436$ in computational linguistics and biological domains [Hirako et al., 2024], whereas Li et al. [2019] record $\rho = 0.556$ when given the additional use of peer review text. Our Qwen2.5-14B model trained on 16% of our corpus obtained a Spearman correlation of $\rho = 0.826$, marking a 27-point improvement.

Scaling-laws and extrapolation. Our empirical findings underscore both the current limitations and future promise of pre-trained language models for citation rate prediction. First, scaling-law analysis reveals a clear, monotonic improvement across all metrics as data volume increases and a similar (albeit more aberrant) increase as model size increases. We hypothesize that the deviation in performance of mid-sized models is due to their specialized fine-tuning, as both Phi-4-mini-instruct and Qwen2.5-Math-7B were fine-tuned prior to our training regime. We believe this accounts for the deviations reported due to the nature of the fine-tuning performed: Phi-4 was fine-tuned for conversation, whereas the Qwen2.5-7B model was fine-tuned for mathematical reasoning, which we would expect to exhibit more forgetful tendencies of prior knowledge than chat fine-tuning. This is support by our findings in Appendix G, where we compared the scaling-laws of DeepSeek’s reasoning distilled (fine-tuned) models to that of their original Qwen2.5 counterparts and found a consistent, albeit minor, drop in performance across data volumes and model sizes.

Gradient saliency analysis. Gradient saliency analysis reveals dominant attribution in title and abstract tokens, with minimal weight attributed to the main content of manuscripts. This is, perhaps, to be expected; titles and abstracts are among the only textual features which are consistently readily available to researchers, not hidden behind paywalls or required downloads. Notably, training without titles/abstracts yields only minor drops in performance ($\Delta r < 0.02$, see Appendix H), indicating these sections serve as convenient surrogates rather than strictly necessary features. This

is optimistic for future work, perhaps suggesting that these models are in fact capable of using the content of the research itself to predict citations, rather than simply reward hacking.

Temporal holdout. Temporal holdout experiments demonstrate the robustness of our approach. Our smallest model maintains performance for up to a month out-of-distribution, and continues to perform better than many previous works up to two years later. The Bloom-560m model was released in 2022, prior to our cutoff date, which verifies that our model’s success has not been due to data leakage during pre-training; these models are not memorizing content, they are learning to extract and quantify highly-citable features in academic texts. The degradation in performance after one month out-of-distribution indicates the potential benefit of an online-learning regime in practical deployments of these models. For more details on the potential effects of pre-training see Appendix I.

5.2 Ethical considerations

The potential misuse of citation-prediction systems is not a subject to be flippantly dismissed. Such tools could facilitate superficial ”citation hacking”, whereby authors tailor linguistic style to game performance metrics, rather than improving the inherent scientific value of their work. Moreover, it is reasonable to expect such tools to find use in publications and journals as a means of automatic sorting, efficiently discriminating between highly citable papers and their less citable counterparts. However, historical precedents should serve to caution against such a system. For example, Alan Turing’s 1936 insights on computability introduced theory that would later become the foundation of modern computer science [Turing et al., 1936]. His work remained almost entirely un-cited for decades and would surely qualify for rejection by an automated citation-prediction based system such as the one we described above, despite its paradigm-shifting effect. It is the belief of the authors that citation-prediction tools should complement, rather than supplant, human judgment, leaving room for the paradigm-shifting works of our time.

5.3 Limitations & future work

Due to limited time and compute, our study was constrained by single-run experiments conducted under a strict seven-day execution window, which inherently restricted the scope and depth of our analysis. This computational budget also necessitated the manual hyperparameter tuning of our framework, which almost certainly resulted in a suboptimal configuration. During this process our models exhibited signs that the initial training phase of the linear head constrained the subsequent full-model fine-tuning to an inferior local minima, a symptom which could likely be mitigated by a more gradual unfreezing of model weights [Howard and Ruder, 2018]. Moreover, our gradient saliency analysis and ablations both suggest an unnecessary reliance on peripheral elements during inference. Finally, our models were inherently limited by their training data, which comprised exclusively of biomedical research written in English.

To address these shortcomings, future work may include:

1. The integration of **multi-modal representations** to more comprehensively model scientific manuscripts.
2. The implementation of an aggressive **section-wise dropout** during training to compel models to recognize weak-but-valuable signals, thereby increasing robustness.
3. Leveraging **progressive training regimes** to prevent trapping the models in suboptimal minima.
4. Expanding evaluations to include **non-biomedical and non-English corpora** to verify the generalizability and real-world applicability of direct LLM-based citation prediction.

6 Conclusion

We have introduced ForeCite, a unified framework for adapting pre-trained causal language models for direct, end-to-end prediction of average monthly citation rates using only in-text information. Through comprehensive scaling-law experiments, we demonstrated that model performance grows

predictably with both parameter count and data volume, and our largest configuration—Qwen2.5-14B on 16% of our 900K article corpus—achieves a Spearman rank correlation of $\rho = 0.826$ on held-out test sets, a 27-point improvement over the previous state-of-the-art. Extrapolations using a bounded least squares model suggest that state-of-the-art LLMs trained on a full dataset could push the performance of our framework significantly farther.

Complementary studies revealed that specialized reasoning fine-tuning does not enhance citation forecasting and may in fact degrade the semantic signals used for this task. Gradient saliency analysis and targeted ablations displayed heavy reliance on titles and abstracts—convenient shortcuts rather than indispensable features—highlighting the need for more intelligent training strategies which encourage full text comprehension.

Altogether, these findings chart a clear path forward: richer multi-modal integration, aggressive dropout schemes to encourage full-text comprehension, and robust evaluation corpora to demonstrate generalizability. While substantial gains remain to be realized at larger scales, ForeCite lays the groundwork for AI-driven research tools capable of guiding academics toward higher-impact writing and, in the long-term, accelerated scientific discovery.

Acknowledgments and Disclosure of Funding

This research was supported by an NSERC Undergraduate Student Research Award (USRA), the NSERC Discovery Grant program and the Canada Research Chairs program. Our deepest thanks to Dr. Shane Arsenault for kindly allowing his work to be featured in our gradient saliency analysis. Our experiments were made possible by the exceptional support team and world-class infrastructure of Compute Canada. We extend a special thanks to Elsevier for granting us access to their publication’s data; without their cooperation, this work could not have been realized. Finally, we thank Peter and Dhiraj for their valuable contributions to early discussions.

Author contributions

The roles of each author are described using the CRediT (Contributor Roles Taxonomy) as follows: **Gavin Hull:** Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing - original draft; Writing - review & editing; Visualization; Project administration; Funding acquisition.

Dr. Alex Bihlo: Conceptualization; Resources; Writing - review & editing; Supervision; Project administration; Funding acquisition.

References

- W. S. N. Aiza, L. Shuib, N. Idris, and N. B. A. Normadhi. Features, techniques and evaluation in predicting articles’ citations: A review from years 2010–2023. *Scientometrics*, 129(1):1–29, 2024.
- S. Arsenault, R. Y. Benoit, F. Clift, and C. S. Moore. Does the use of the bruton tyrosine kinase inhibitors and the c-kit inhibitor masitinib result in clinically significant outcomes among patients with various forms of multiple sclerosis? *Multiple Sclerosis and Related Disorders*, 67:104164, 2022.
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM joint conference on digital libraries*, pages 351–360. IEEE, 2014.
- F. Davletov, A. S. Aydin, and A. Cakmak. High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 491–498, 2014.

- T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL <https://arxiv.org/abs/2305.14314>.
- Y. Hao, L. Dong, F. Wei, and K. Xu. Self-attention attribution: Interpreting information interactions inside transformer, 2021. URL <https://arxiv.org/abs/2004.11207>.
- J. Hirako, R. Sasano, and K. Takeda. Realistic citation count prediction task for newly published papers. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1131–1141, Dubrovnik, Croatia, may 2023. doi: 10.18653/v1/2023.findings-eacl.84. URL <https://aclanthology.org/2023.findings-eacl.84/>.
- J. Hirako, R. Sasano, and K. Takeda. Cimate: Citation count prediction effectively leveraging the main text, 2024. URL <https://arxiv.org/abs/2410.04404>.
- J. Howard and S. Ruder. Universal language model fine-tuning for text classification, 2018. URL <https://arxiv.org/abs/1801.06146>.
- A. Ibáñez, P. Larrañaga, and C. Bielza. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309, 2009.
- N. Jamal, M. Alauthman, M. Malhis, and A. M. Ishtaiwi. Forecasting research influence: a recurrent neural network approach to citation prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, 36(2):1070–1082, 2024.
- J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.
- J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in nlp, 2016. URL <https://arxiv.org/abs/1506.01066>.
- S. Li, W. X. Zhao, E. J. Yin, and J.-R. Wen. A neural citation count prediction model based on peer review text. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4914–4924, Hong Kong, China, nov 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1497. URL <https://aclanthology.org/D19-1497/>.
- N. Pobiedina and R. Ichise. Citation count prediction as a link prediction problem. *Applied Intelligence*, 44:252–268, 2016.
- X. Ruan, Y. Zhu, J. Li, and Y. Cheng. Predicting the citation counts of individual papers via a bp neural network. *Journal of Informetrics*, 14(3):101039, 2020.
- A. M. Turing et al. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.
- T. van Dongen, G. Maillette de Buy Wenniger, and L. Schomaker. Schubert: Scholarly document chunks with bert-encoding boost citation count prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, page 148–157. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.sdp-1.17. URL <http://dx.doi.org/10.18653/v1/2020.sdp-1.17>.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.
- D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, Oct. 2013. ISSN 1095-9203. doi: 10.1126/science.1237825. URL <http://dx.doi.org/10.1126/science.1237825>.
- G. M. d. B. Wenniger, T. van Dongen, and L. Schomaker. Multischubert: Effective multimodal fusion for scholarly document quality prediction. *arXiv preprint arXiv:2308.07971*, 2023.

A Hyperparameters

Hyperparameters		
Group / Parameter		Value
Training	Optimizer	AdamW
	Learning Rate	1.0e-4
	Learning Rate Scheduler	Cosine
	Weight Decay	1.0e-2
	Grad. Accum. Steps	4
	Batch Size	2
	Epochs	1
Fine-tuning	Optimizer	AdamW
	Learning Rate	1.0e-4
	Learning Rate Scheduler	Cosine
	Weight Decay	1.0e-2
	Grad. Accum. Steps	16
	Batch Size	1
	Epochs	3
QLoRA	Dropout	5.0e-2
	Alpha	8
	Rank	4

Figure 6: A table of hyperparameters used for all experiments. Any omitted parameters were left at their default values.

B Model variants

Model	Publisher	Parameters (B)
Bloom-560m	BigScience	0.56
Llama-3.2-1B	Meta	1.24
Phi-4-mini-instruct	Microsoft	3.84
Qwen2.5-Math-7B	Alibaba	7.62
Qwen2.5-14B	Alibaba	14.8
DeepSeek-R1-Distill-Qwen-7B	DeepSeek	7.62
DeepSeek-R1-Distill-Qwen-14B	DeepSeek	14.8

Figure 7: This table contains every base language model used throughout our research, as well as their parameter counts and publishers.

C Evaluation metrics

To evaluate model performance, we reported the following metrics:

- **Pearson correlation coefficient (r):** Measures the linear relationship between predicted and true (transformed) citation rates. This is our most commonly reported statistic due to its ubiquity and simplicity.
- **Spearman rank correlation (ρ):** Captures the monotonic relationship between predicted and true (transformed) citation rates. Equivalent to the Pearson correlation between the rank values of the two variables. This is the statistic we most commonly compare to prior works, as it is invariant under any monotonic transformation, ensuring compatibility between reportings.
- **Coefficient of determination (R^2):** Quantifies the proportion of the variation in the true (transformed) citation rates which is explainable by the predicted citation rates, and thus our models.
- **Mean absolute error (MAE):** Represents the average magnitude of the prediction error.
- **Mean squared error (MSE):** Similar to MAE, except it penalizes large deviations more heavily.

D Scaling-laws

Scaling-laws								
Model name	Parameters	Split	Data %	r (\uparrow)	ρ (\uparrow)	R^2 (\uparrow)	MSE (\downarrow)	MAE (\downarrow)
Bloom-560m	0.559	train	1	0.502	0.473	0.251	0.749	0.677
			2	0.565	0.539	0.318	0.682	0.645
			4	0.660	0.633	0.433	0.567	0.588
			8	0.704	0.673	0.494	0.506	0.556
			16	0.749	0.716	0.560	0.440	0.519
		test	1	0.430	0.380	0.181	1.342	0.911
			2	0.561	0.546	0.313	1.064	0.795
			4	0.609	0.579	0.367	0.992	0.770
			8	0.680	0.652	0.459	0.841	0.714
			16	0.721	0.685	0.518	0.745	0.674
Llama-3.2-1B	1.24	train	1	0.719	0.688	0.516	0.484	0.544
			2	0.794	0.765	0.630	0.370	0.476
			4	0.824	0.800	0.677	0.323	0.443
			8	0.848	0.824	0.718	0.282	0.414
			16	0.862	0.839	0.742	0.258	0.396
		test	1	0.674	0.619	0.453	0.902	0.740
			2	0.720	0.682	0.511	0.702	0.652
			4	0.741	0.717	0.544	0.665	0.637
			8	0.777	0.742	0.599	0.609	0.612
			16	0.787	0.758	0.615	0.596	0.603
Phi-4-mini-instruct	3.84	train	1	0.763	0.738	0.582	0.418	0.506
			2	0.804	0.774	0.646	0.354	0.465
			4	0.843	0.817	0.711	0.289	0.421
			8	0.865	0.844	0.747	0.253	0.391
			16	0.872	0.852	0.760	0.240	0.381
		test	1	0.657	0.630	0.423	0.876	0.734
			2	0.723	0.698	0.519	0.727	0.659
			4	0.752	0.727	0.559	0.623	0.613
			8	0.789	0.767	0.619	0.578	0.592
			16	0.802	0.781	0.641	0.549	0.575
Qwen2.5-Math-7B	7.62	train	1	0.759	0.728	0.575	0.425	0.507
			2	0.810	0.780	0.656	0.344	0.459
			4	0.852	0.826	0.725	0.275	0.410
			8	0.873	0.851	0.761	0.239	0.380
			16	0.890	0.870	0.790	0.210	0.356
		test	1	0.645	0.607	0.411	0.880	0.736
			2	0.721	0.696	0.517	0.752	0.678
			4	0.753	0.724	0.563	0.648	0.626
			8	0.776	0.750	0.598	0.625	0.615
			16	0.786	0.758	0.613	0.600	0.602
Qwen2.5-14B	14.8	train	1	0.811	0.791	0.654	0.346	0.454
			2	0.866	0.848	0.746	0.254	0.388
			4	0.912	0.898	0.831	0.169	0.316
			8	0.924	0.911	0.852	0.148	0.295
			16	0.936	0.924	0.875	0.125	0.270
		test	1	0.752	0.754	0.561	0.653	0.615
			2	0.797	0.779	0.632	0.544	0.575
			4	0.826	0.797	0.679	0.487	0.544
			8	0.832	0.809	0.687	0.472	0.534
			16	0.844	0.826	0.706	0.452	0.522

Figure 8: Results of scaling-law analysis. Model sizes span 0.5B to 14B parameters and data volumes span between 1% ($\approx 90K$) and 16% ($\approx 144K$). Best performance is found in **bold** for both train and test splits.

E Scaling-law extrapolation

Least squares fit parameters			
	β_0	β_1	β_2
$r(p, d)$	0.6771	0.0689	0.0767
$\rho(p, d)$	0.6260	0.0698	0.0724

Figure 9: Model parameters derived for scaling-laws using least squares.

F Gradient saliency heatmap

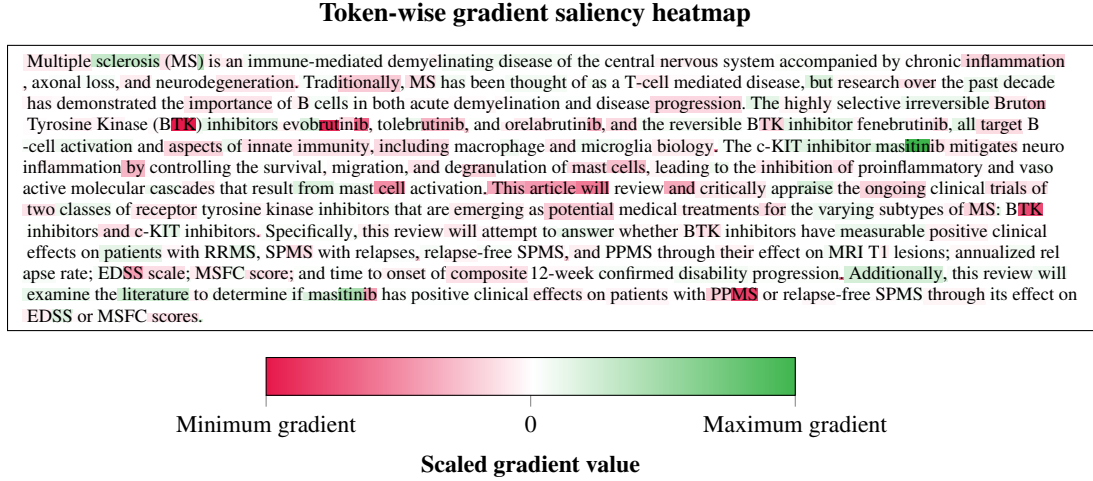


Figure 10: A token-level heatmap of model gradients for Arsenault et al. [2022], computed using Llama-3.2-1B trained on 16% of our corpus. Red tokens decreased predicted citation rate, whereas green tokens increased the predicted citation rate. Color intensity is proportional to gradient magnitude, reflecting intensity of impact on predicted citation rate.

G Large reasoning models

We examined the effect of specialized reasoning fine-tuning on citation prediction by comparing each Qwen2.5 checkpoint to its DeepSeek distilled counterpart. Figure 11 plots Pearson correlation for the original and distilled Qwen2.5 models at increasing data fractions. At each evaluation of 4%+ data volume, both Qwen2.5 variants surpassed their DeepSeek counterparts, indicating that reasoning fine-tuning does not improve, and may even slightly reduce, citation prediction performance.

This suggests that the latent semantic knowledge stored inside the LLMs which is most useful during citation prediction is contextual information rather than logical information, as we would expect contextual information accumulated during training to deteriorate during fine-tuning of any kind. This theory is consistent with the deviation in performance reported from the Phi-4-mini-instruct and Qwen2.5-Math-7B models during scaling-law analysis. We suspect this trend may reverse as more "intelligent" models begin to use the main content of articles for their predictions rather than relying on title and abstract tokens.

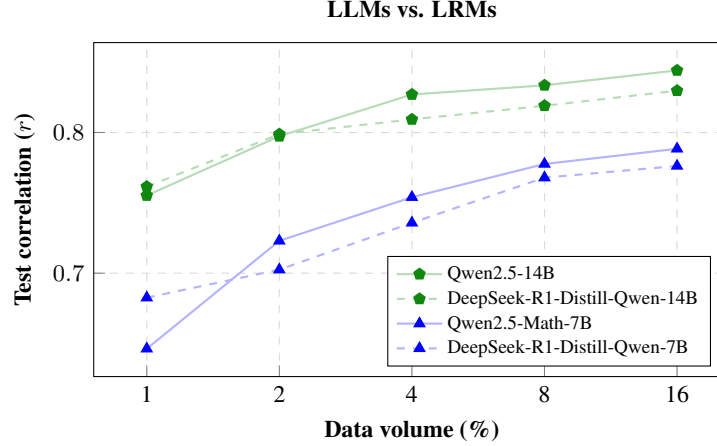


Figure 11: Effect of reasoning fine-tuning on citation prediction performance as measured in Pearson correlation at varying degrees of data availability.

H Ablations

To assess the aforementioned reliance on peripheral elements in sample articles, we performed two targeted ablation studies on Llama-3.2-1B (16% data): in one we removed all title tokens and in the other we removed all abstract tokens.

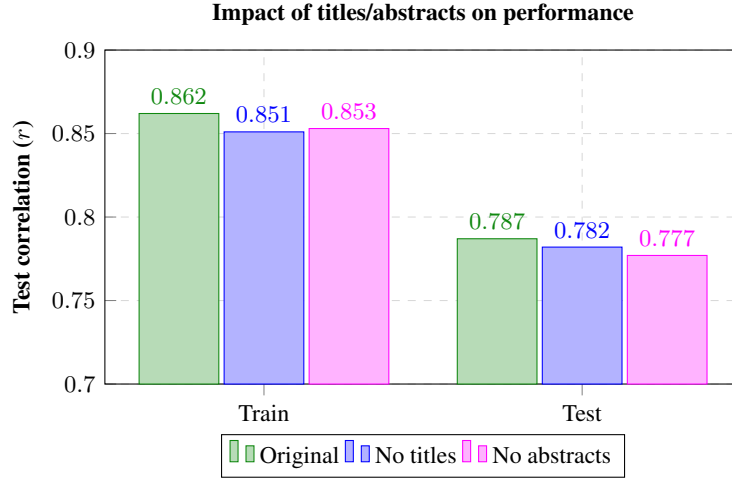


Figure 12: Pearson correlation for full text vs. no-title vs. no-abstract ablations.

Figure 12 displays the results: the full model achieved $r = 0.787$, the no-title variant $r = 0.782$, and the no-abstract variant $r = 0.777$. These relatively negligible drops suggest that the prior reliance on abstract and title tokens may have been a reflection of convenience rather than necessity. This may motivate the use of section-by-section dropout schemes in future work to encourage more robust predictions.

I Effect of pre-training

In order to verify the utility of pre-training in our predictions, we evaluated Llama-3.2-1B on various data volumes after pre-training (the original model), and after a random re-initialization (simulating the model prior to pre-training). Figure I demonstrates a steep drop in performance after re-initialization across all data volumes. The observed performance drop narrowed from $\Delta r = 0.617$ to $\Delta r = 0.313$ as data volume increased.

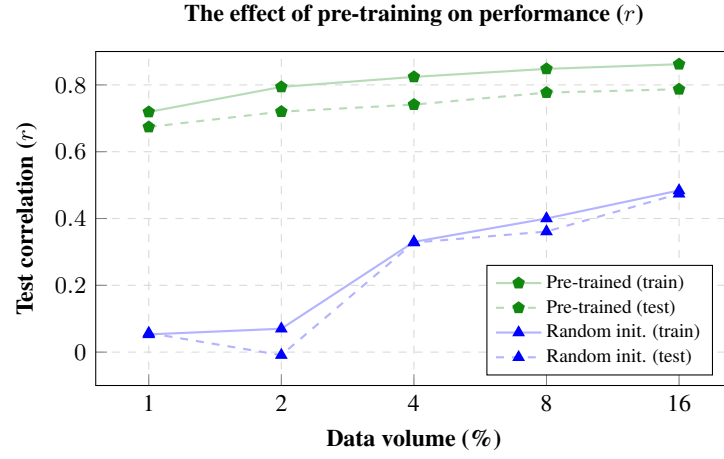


Figure 13: Effect of pre-training on citation prediction performance as measured in Pearson correlation at varying degrees of data availability.

These results confirm our suspicions that the semantic knowledge learned during pre-training is useful in discerning the features which characterize highly-citable scientific text.