Opening Lines that Matter: Novel Abstracts and their Citations

Lalit Chaudhary [*] **Urmi Jana** [*] **Edith Meinzinger** [*]
Kanchan Poudel [*] **Umer Shahzad** [*]

Abstract

Citation counts are widely used as proxies for research impact, yet the factors driving them remain incompletely understood. In this study, we investigate whether the semantic novelty of a paper's title and abstract is associated with its citation impact. We analyse 2.5 million papers across three domains: artificial intelligence, physics, and psychology published since 2000. Novelty is quantified using a semantic embedding space and measured relative to each paper's nearest neighbors. Our results indicate that the relationship between novelty and citation impact is strongly discipline-dependent. These findings highlight the role of field-specific norms in shaping how novel research is received and cited. We include the source code and the processed dataset at github.com/lalit3c/DataLit-Scientific-Novelty-Impact

1. Introduction

Publish or perish is a highly debated phrase in research. More often than not, citation counts and h-indices are referenced as metrics for the relevance of a scientist's work and "quality" of research. Authors involved in scientific publications spend a significant amount of time ensuring their writing is clear and widely read. A catchy title surely captures the attention of the reader, but the abstract should convince the reader of the relevance the study brings to the field. We assume that an abstract introducing novel ideas significantly increases the chances of the paper being read. But does a novel abstract also impact the citation count?

There certainly will be some skepticism about groundbreaking new results. Researchers might have a feeling of results being to good to be true, but on the other hand, redundant papers do not receive much attention in the scientific community. It might be difficult to find the right balance between novelty and conventionality, and furthermore, we expect that such a balance might depend on the field of research.

In this work, we study the novelty of a paper using textual embeddings. We primarily focus on the following questions:

(1) Based solely on the title and abstract, do we see a relationship between the novelty of a paper and its citation count? (2) Is there a difference in the reception of novel papers between different scientific fields?

2. Data and Methods

Since we aim to analyse novelty across varying scientific fields, we focus on three distinct domains: artificial intelligence (AI), psychology, and physics. Figure 1 shows our methodological procedure.

Preprocessing the data

The dataset was constructed by integrating bibliographic records from two large-scale open-access databases: OpenAlex (Priem et al., 2022) and Semantic Scholar (Kinney et al., 2025). OpenAlex provides structured disciplinary classifications for each publication, including field, primary topic, and subtopic assignments. Semantic Scholar was used as a reliable source of abstracts and citation counts.

From Semantic Scholar, we extracted core metadata for each paper, including corpus ID, title, publication date, citation count, and field of study. To avoid downloading the entire terabytes worth of Semantic Scholar corpus, we performed on-the-fly gzip decompression using Python's zlib library, processing data in streaming chunks and filtering records in real time to retain only publications relevant to the selected domains. An analogous dataset was constructed from OpenAlex to validate field assignments and to obtain more fine-grained topic classifications.

The metadata table was subsequently augmented with full-text abstracts retrieved from another Semantic Scholar dataset. For each record, we extracted the corpus ID, abstract text, and available external identifiers (DOIs, arXiv IDs, and PubMed IDs), enabling cross-database reconciliation. The datasets from the two sources were then merged

[*]Equal contribution . Correspondence to: EM <edith.meinzinger@student.uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

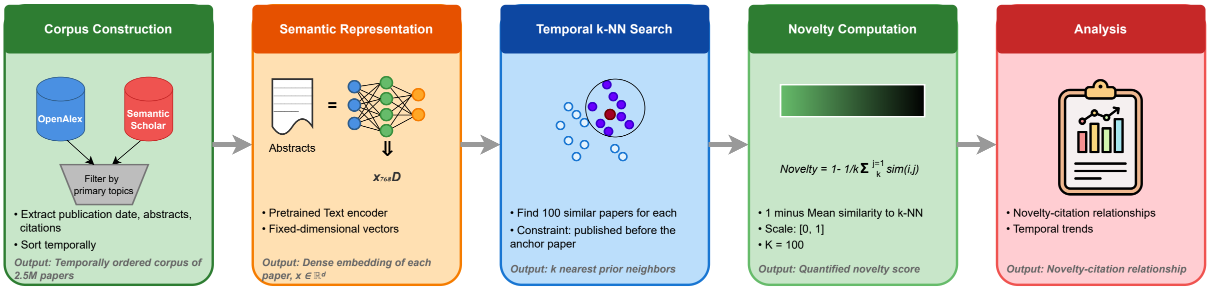

Figure 1. Overview of the data collection, processing, and analysis pipeline.

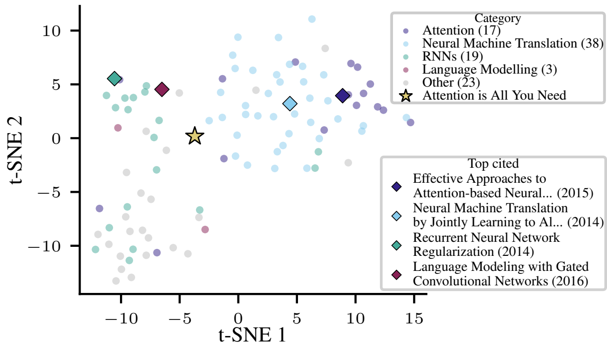

Figure 2. t-SNE visualization of the local embedding neighborhood among temporal predecessor set of "Attention Is All You Need" (star). Each point represents a paper embedding, colored by category; diamond markers indicate the most cited papers from each category.

and cleaned. Record linkage was performed using a hierarchical multi-key matching strategy executed in the following priority order: (1) exact title matching, (2) DOI matching and (3) Microsoft Academic Graph (MAG) ID matching extracted from Semantic Scholar's external identifiers. To ensure temporal consistency and metadata completeness, we retained only papers published between January 2000 and December 2025 with complete metadata fields.

The resulting corpus comprises 2,516,624 publications, including 877,871 papers in psychology, 955,982 in physics, and 725,942 in artificial intelligence. [1]

Quantifying novelty

Novelty as a predictor for the impact of a paper has been studied by several others before with a focus on on bibliometric analysis or citation graphs. Uzzi et al. (2013) conceptualized novelty by analysing the references of a paper and evaluating the novel pairings of papers or journals. With the rise of large language models, recent works involve text-based approaches with a focus on keywords and sentence-based concepts (Yan et al., 2019; Chen & Fang, 2019). However, the novelty of the abstract has not been considered as a significant factor in prior works. The analysis of the abstract follows this idea and matches the actual reading process of researchers of new studies.

To compute the novelty, we constructed an embedding space of all papers using the SPECTER2 model which is pretrained to genereate embeddings for scientific tasks (Cohan et al., 2020). For each paper, the hundred approximately nearest neighbors were identified, excluding any papers published after the current paper's publication month (Figure 2). Approximate nearest neighbor search was performed using FAISS (Johnson et al., 2019) to ensure scalability across the full corpus. To ensure that each paper has a meaningful set of prior neighbors, the novelty metric was only computed for papers published after January 2005. This idea is formulated below.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denote the corpus of papers, where each paper p_i is associated with a publication date t_i and an embedding $\mathbf{e}_i \in \mathbb{R}^{768}$. For a given paper p_i, we define its temporal predecessor set as:

$$\mathcal{P}_i^{<} = \{p_j \in \mathcal{P} : t_j < t_i\}. \tag{1}$$

Let $\mathcal{N}_k(p_i) \subseteq \mathcal{P}_i^{<}$ denote the set of its $k = 100$ nearest neighbors in embedding space. The novelty score of paper p_i is defined as:

$$\text{Novelty}(p_i) = 1 - \frac{1}{k} \sum_{p_j \in \mathcal{N}_k(p_i)} \text{sim}(p_i, p_j), \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between embeddings.

3. Results

Since the range of raw novelty scores was quite small (0 to 0.216), we normalized the scroes to 0 to 1 range for our downstream analysis. We summarize our findings in Figure 3.

[1]The dataset is available at https://huggingface.co/datasets/lalit3c/S2_CS_PHY_PYSCH_papers

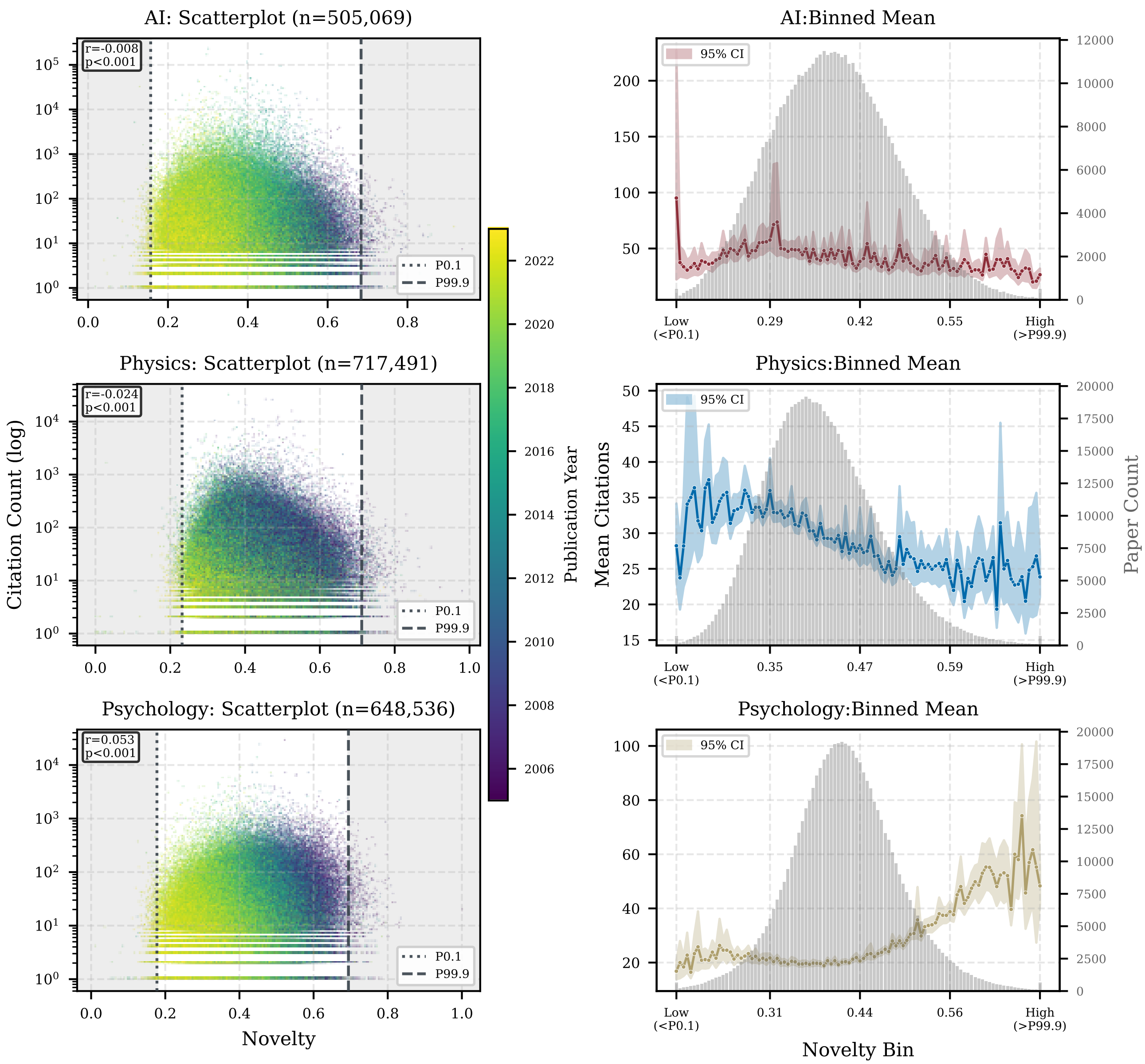

Figure 3. Relationship between novelty and citation count across three scientific disciplines. (**Left column**) Scatter plots showing individual papers' citation counts (log scale) versus novelty scores, coloured by publication year. Dotted and dashed vertical lines indicate field-specific percentile cutoffs (0.1st and 99.9th percentiles, respectively). Pearson correlation coefficients (r) and associated p-values are displayed for each field. (**Right column**) Mean citation counts computed within 100 fixed-width novelty bins. For each bin, the mean novelty and mean citation are plotted, with 95% bootstrapped confidence intervals shown as shaded bands. Grey bars indicate the number of papers per bin. Edge bins below 0.1st (Low) and above 99th (High) percentile are merged to account for sparse data.

We observe a clear temporal trend across all fields: older papers tend to show higher novelty than more recent publications. This effect is particularly pronounced in psychology and AI, whereas it is less so in physics.

To further analyse the relationship between novelty and impact, we discretized the novelty scores within each field into a hundred fixed-width bins and computed the mean citation count for each bin. Given the highly skewed distribution of citations, we estimated 95% confidence intervals for the mean using 1,000 bootstrap samples to quantify uncertainty. While the number of papers per bin follows an approximately Gaussian distribution across all fields, the relationship between novelty and mean citations indicate distinct patterns in different disciplines.

In AI, no clear relationship is observed between novelty and mean citation count, and the confidence intervals likewise indicate minimal effects at the extremes of novelty. In physics, papers with lower novelty tend to receive more citations on average. The confidence intervals for both low- and high-novelty bins are wider, reflecting greater uncertainty due to the smaller number of papers in these ranges. Psychology exhibits a markedly different pattern: papers with low to average novelty have the lowest mean citation counts with citations increasing alongside novelty. Similar to physics, higher-novelty bins show wider confidence intervals, reflecting fewer papers in these ranges and consequently greater uncertainty.

4. Discussion

Our analysis reveals distinct relationships between novelty and citation across scientific fields. In physics, papers with moderate novelty tend to receive higher citations. This pattern likely reflects the field's reliance on established theoretical frameworks and rigorous mathematical evidence, where extremely novel contributions must provide strong, reproducible results to be accepted. Consistent with Uzzi et al. (2013), papers that combine novelty with conventionality often achieve the greatest impact, which may explain why lower-novelty papers in physics receive higher citations in our dataset.

In contrast, psychology, as a relatively young field with weaker theoretical constraints (Muthukrishna & Henrich, 2019), exhibits a trend in which higher novelty is generally associated with greater citations. The "Feeling the Future" case (Bem, 2011) illustrates how striking claims can achieve publication despite replication challenges. At last, the replication crisis further highlights structural factors that can favor surprising or novel findings. Wiggins & Christopherson (2019) provides an overview of research practices that may promote novelty and its visibility.

AI shows no consistent overall relationship between novelty and citations. The peaks observed in Figure 3 are likely driven by seminal, highly cited papers such as "Attention Is All You Need" (Vaswani et al., 2017), which has a novelty score of 0.29 and contributes to a local peak in mean citations. More broadly, the lack of correlation may result from the field's recent focus on a few dominant topics, generating a large volume of related papers with similar citation counts and consequently low relative novelty.

Limitations

Our approach assumes that a paper's abstract sufficiently captures its novel contributions. Moreover, the proposed novelty metric depends on the quality of textual embeddings and on the extent to which they encode conceptual, rather than purely semantic, differences between papers. These limitations are illustrated by "Attention Is All You Need" (Vaswani et al., 2017). Despite its seminal nature, the empirical novelty score is 0.29, significantly lower than most papers in our AI dataset. This may be because the abstract primarily discusses attention and machine translation, and its nearest neighbors already address similar concepts in earlier work, as shown in Figure 2. Similarly, the metric is based on embeddings of papers published from 2000 onward and is temporally constrained to search only past neighbors. While this constraint is necessary to assess novelty relative to prior work, earlier papers may receive inflated novelty scores due to fewer comparable neighbors, whereas increasing density in the embedding space leads to systematically lower novelty scores for more recent papers.

5. Conclusion

Our findings suggest that semantic novelty in abstracts is not a universal predictor of citation impact, but instead interacts strongly with disciplinary norms. While novelty may enhance visibility and influence in some fields, it can be neutral or even detrimental in others, reflecting differences in how communities evaluate and adopt new ideas. These results emphasize the importance of considering domain-specific conventions and contribute to a more nuanced understanding of how scientific communication impacts the respective community.

Contribution Statement

In order of appearance in the author list:

Lalit Chaudhary scoped, collected and prepared the dataset, constructed the embeddings for the abstracts and provided visualizations for the results.

Urmi Jana scoped, collected and prepared the dataset, provided visualizations for the t-sne representations and added to the GitHub repository.

Edith Meinzinger provided the literature review and identified potential gaps in the existing work, collected the data and performed the analysis of the papers in psychology.

Umer Shahzad collected, prepared the data and performed analysis for papers in physics.

Kanchan Poudel scoped, collected and prepared the dataset and constructed the novelty score we use in our work, along with the visual abstract included in the report.

All authors jointly wrote the report.

References

Bem, D. J. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3): 407–425, 1 2011. doi: 10.1037/a0021524. URL https://doi.org/10.1037/a0021524.

Chen, L. and Fang, H. An automatic method for extracting innovative ideas based on the Scopus® database. *KNOWLEDGE ORGANIZATION*, 46(3):171–186, 1 2019. doi: 10.5771/0943-7444-2019-3-171. URL https://doi.org/10.5771/0943-7444-2019-3-171.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. SPECTER: Document-level Representation Learning using Citation-informed Transformers. *ArXiv*, abs/2004.07180, 2020. URL https://api.semanticscholar.org/CorpusID:215768677.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S., Gorney, J., Graham, D., Hu, F., Huff, R., King, D., Kohlmeier, S., Kuehl, B., Langan, M., Lin, D., Liu, H., Lo, K., Lochner, J., MacMillan, K., Murray, T., Newell, C., Rao, S., Rohatgi, S., Sayre, P., Shen, Z., Singh, A., Soldaini, L., Subramanian, S., Tanaka, A., Wade, A. D., Wagner, L., Wang, L. L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., Zuylen, M. V., and Weld, D. S. The semantic scholar open data platform, 2025. URL https://arxiv.org/abs/2301.10140.

Muthukrishna, M. and Henrich, J. A problem in theory. *Nature Human Behaviour*, 3(3):221–229, 2019. doi: 10.1038/s41562-018-0522-1. URL https://doi.org/10.1038/s41562-018-0522-1.

Priem, J., Piwowar, H., and Orr, R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. URL https://arxiv.org/abs/2205.01833.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 10 2013. doi: 10.1126/science.1240474. URL https://doi.org/10.1126/science.1240474.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wiggins, B. J. and Christopherson, C. D. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39(4):202, 2019. doi: doi.org/10.1037/teo0000137. URL https://doi.org/10.1037/teo0000137.

Yan, Y., Tian, S., and Zhang, J. The impact of a paper's new combinations and new components on its citation. *Scientometrics*, 122(2):895–913, 12 2019. doi: 10.1007/s11192-019-03314-6. URL https://doi.org/10.1007/s11192-019-03314-6.