Anticipating Retractions in Scientific Databases using LLM-Based Citation Analysis

Muhammad Usman[0000-0002-6154-6256] ✉, Mayukh Das, and Wolf-Tilo Balke[0000-0002-5443-1215]

Institute for Information Systems, TU Braunschweig, Braunschweig, Germany
{Usman,mayukh,Balke}@ifis.cs.tu-bs.de

Abstract. The increasing number of retracted articles raises concerns about scientific reliability, as they can spread flawed information. Moreover, uninformed and pre-retraction citations of these articles pose a cascading threat to scientific integrity. With no systematic method to identify articles at risk of retraction, we aim to address this challenge by detecting those susceptible to retraction and requiring further evaluation. We propose a triage process that focuses on Concerning Citations (CCs)—citations in which the citing paper questions the validity of a cited study's data, methodology, or conclusions. To establish a foundation for this process, we developed a dedicated dataset to detect CCs, creating a new classification task for AI-based early identification of articles at risk of retraction. We evaluated machine learning (ML), encoder-based, and decoder-based large language models (LLMs) on this task, investigating the impact of scale, supervised fine-tuning (SFT), and few-shot learning on model performance. Our findings indicate that encoder-based models, particularly BERT, outperform other models. While scale and few-shot learning benefit large decoder models (e.g., LLaMA2), SFT does not consistently improve performance. This study contributes to the early identification of potential retractions, helping mitigate their impact on the scientific community.

Keywords: Citations Analysis, · Retraction Analysis.

1 Introduction

The scientific community relies heavily on published research to progress across various disciplines. In this context, scientific databases are crucial as repositories and facilitators of scholarly communication. The efficacy of these databases in supporting knowledge advancement depends on their capacity to maintain the accuracy of published research. However, the growing number of retracted articles in recent years has raised significant concerns within the scientific community [19]. Articles can be retracted for various reasons, such as falsification, fabrication, or plagiarism (FFP). Retracted articles undermine the credibility of published research and perpetuate flawed information when cited, either before retraction or unknowingly after their retraction. However, continuously re-evaluating all peer-reviewed articles for unreliable information—whether due to intentional deceit or honest errors—is impractical.

Citations are essential for reviewing scientific literature, identifying research gaps, discussing the strengths and limitations of existing studies, and advancing scholarly dialogue [5]. However, their role in highlighting potential concerns in existing articles is often overlooked. These citations can highlight problematic research early, serving as a crucial mechanism for preemptively identifying articles with unreliable information. For example, in 2006, Erler et al. [6] suggested a role for HIF2 in regulating the LOX gene. In 2014, Xu et al. [21] raised concerns, noting that while functional studies had identified the binding site of HIF1, they did not address the potential direct interaction of HIF2. Xu et al. questioned whether HIF2 directly bound to the LOX gene, stating that "although functional studies have successfully identified the binding site in the LOX gene for HIF1 [r40], whether there is a direct binding effect of HIF2 to the LOX gene remains unknown." Notably, six years later, In March 2020, the Erler et al. study was retracted due to lack of data to validate the findings [7], validating the original concerns raised in the citation. Addressing such concerns earlier would help scrutinize articles containing unreliable information and prevent the perpetuation of false findings. Through a comprehensive examination of cases in which experts have identified concerns regarding existing research (i.e., concerning citations), we propose a robust self-correction mechanism for the scientific community. This approach enables the identification of potentially unreliable research without relying on external investigations (see Fig. 1). Given the limitations of existing research, the key contributions of this work are as follows.

- **Citations as Early Warning Signs:** We investigate how citations can help identify articles that require re-evaluation and potential retraction.
- **Benchmarking AI Models for Detecting Concerning Citations:** We evaluate the performance of various machine learning and large language models in detecting concerning citations, analyzing the effectiveness of different approaches, including the impact of scale, supervised fine-tuning, and few-shot learning.

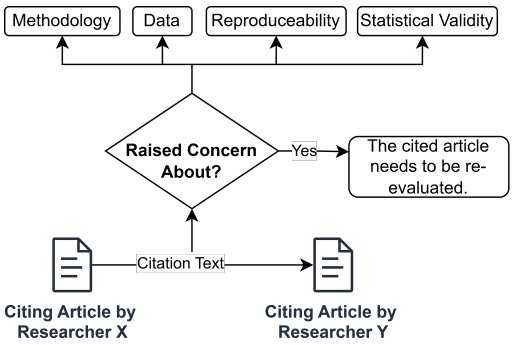

Fig. 1: Potential concerns that can be raised in citation text.

2 Literature Review

Scientific research is not merely a quest for knowledge and facts but also a reflection of human behavior, encompassing brilliance, errors, and misconduct. The increasing incidence of retracted articles raises significant concerns about the reliability of published research [19]. Due to the subjective nature of retraction, existing literature primarily focuses on its quantitative aspects, including the rate of retraction [15], time of retraction [9]. Moreover, despite being indexed by databases such as PubMed, Web of Science, and Scopus, retracted articles are often cited without acknowledgment of their retraction status, undermining the accuracy of subsequent studies [1].

Table 1: Examples of concerning citations that highlight issues in methodology, data quality, statistical validity, sample size, or reproducibility.

Definition of Concerning Citation
Concerning citations occur when a citing paper questions the validity of a cited study, addressing flaws in data, methodology, or conclusions.
Examples
Methodological Concerns: Smith and Clark (2017) used a novel approach, but ignored confounding variables, affecting validity. ***Data Quality Concerns:*** Jones et al. (2015) reported significant correlation, but their dataset had inconsistencies. ***Statistical Validity Concerns:*** Lee and Kim (2018) reported significant p-values, but lacked correction for multiple comparisons. ***Sample Size Concerns:*** Brown and Davis (2016) made predictions on a small dataset, limiting reliability. ***Reproducibility Concerns:*** Garcia et al. (2019) introduced an algorithm, but replications failed to yield similar results.

2.1 Citation As Quality Control

Citations are fundamental to scientific communication. Researchers have identified over 150 citation intention categories [11]. Despite advancements, accurately identifying citation intentions remains a challenging task due to its nuanced nature, underscoring the need for a specialized approach [12]. Recent studies have classified citations into categories such as important and non-important [18], as well as dependent and non-dependent [17]. Despite extensive discussions on citation intentions, their potential for quality control and error identification in scientific literature remains underexplored. Te et al. classify citations as critical and non-critical [16] based on Bordignon's definition of critical citations [2]. However, this framework does not directly support quality control. For example, the 'compare' function, illustrated by statements like 'Y1 and Y2 (2008) outperformed Y3 and Y4 (2007),' often highlights the superiority of newer studies without pointing out flaws in the outperformed ones.

To explore the potential of citations as a quality control mechanism, we define two categories: Concerning Citations (CCs) and Non-Concerning Citations

(NCCs). CCs refer to instances where the citing paper raises concerns about the validity of a cited study due to issues with data, methodology, or other potential flaws (see Table 1). In contrast, NCCs include citations that discuss or criticize a study without questioning its overall validity. This categorization aims to enhance quality control by effectively identifying studies with potential concerns, thereby supporting error identification in scientific literature. To validate our approach, we conducted a retrospective analysis of the ten most-cited retracted articles, investigating their pre-retraction citations for CCs. This analysis aims to determine whether our categorization accurately identifies citations that contributed to subsequent retractions.

3 Corpus of Concerning Citations

Bordignon et al. (2024) published a corpus of 505 critical citations [3], from which we filtered those raising concerns about the cited work. We initially focused on keywords such as "doubt," "question," or "flaw" to identify concerning citations, although these terms can also appear in general critiques. To address potential ambiguity, we manually annotated the citations according to the annotation guidelines available on the GitHub public repository. We identified 63 concerning citations. The remaining 442 citations in the Bordignon et al. (2024) corpus were classified as non-concerning.

Retrospective Analysis of the Top 10 Most-Cited Retracted Articles: Retrospective analysis examines past events or data to draw conclusions about their impact. Commonly used in fields such as healthcare, business, and social sciences, it involves collecting historical data, applying analysis, and interpreting results to guide future decisions [13]. This retrospective analysis of the top 10 highly cited retracted articles[1] examines the presence of concerning citations in the pre-retraction phase, suggesting that these citations could help identify research needing further investigation or potential retraction in the future. We first collected 3,215 pre-retraction citing articles from PubMed Central. Using the Hsiao et al. dataset [10], which includes 2 million open-access articles organized as sentences with in-text citation PMCIDs, we matched citation sentences with the PubMed IDs of retracted articles. By considering one sentence before and after each match, we compiled 4,034 citation contexts. Following the method described in Section 3.1, we identified 56 concerning citations, with eight of the ten retracted articles having more than one concerning citation. For example, the article PubMed ID: 22088800, retracted in 2019 after being published in 2011, took eight years to be retracted. However, citations PubMed ID: 25460005 (2014) and PubMed ID: 28979901 (2016) raised concerns about the study's data years earlier. Despite these concerning citations, retraction was delayed due to insufficient scrutiny. It can be reasonably assumed that addressing these concerns earlier would have resulted in a reduced time to retraction. By flagging

[1] https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/

and promptly investigating concerning citations during peer review, the proposed approach could reduce the time needed to identify dubious findings and help maintain the reliability of scientific literature.

Scaling Concerning Citations with Sci-BERT We identified 119 concerning citation instances by filtering the Bordignon (2024) dataset and conducting a retrospective analysis. However, manual annotation is not scalable. To improve efficiency, we aim to fine-tune Sci-BERT, a language model trained on scientific texts, using this dataset. Our goal is to assess its ability to identify citation patterns linked to retracted articles from 1,005 pre-retraction citations by Usman and Balke (2023) [17]. To annotate unlabeled citations, we fine-tuned the Sci-BERT model using a manually labeled dataset of 269 instances, including 119 concerning citations. To ensure unbiased fine-tuning, we randomly selected 150 non-concerning instances from Bordignon's corpus. The dataset was split into training (80%) and validation (20%) sets. We pre-processed the text with the Sci-BERT tokenizer, applying padding and truncation to a length of 512 tokens, and defined training parameters with a batch size of 8 and 3 epochs. After training, the model was evaluated on the validation set, achieving a precision of 0.85, recall of 0.83, and an F1 score of 0.84. The trained model then predicted labels for 1,005 unlabeled instances, identifying 59 as concerning citations. Manual validation revealed 12 false positives that only highlighted limitations without questioning core validity. Sci-BERT correctly labeled 47 concerning citations. In total, we identified 166 concerning citations. In the subsequent section, we evaluate the performance of different models in classifying citations as concerning or non-concerning and assess their suitability for this critical task.

4 Classification and Evaluation

We now evaluate the efficacy of conventional machine learning and deep learning models on our dataset of concerning and non-concerning citations, focusing specifically on recent advancements in large language models (LLMs). For systematic comparison, we categorized models by architecture type: encoder-based, decoder-based, and encoder-decoder-based. This categorization clarifies structural differences and allows us to assess each architecture's contribution to task performance. Given the dataset's skew toward non-concerning citations, we used weighted metrics—precision, recall, and F1-score.

Machine Learning Models We evaluated the performance of state-of-the-art machine learning models in identifying Concerning Citations. First, we used regular expressions to remove reference markers in various formats (e.g., "[12]," "(12)," "(Author et al.)") from the citation text, while manually excluding non-standard formats. For feature extraction, we applied bi-grams to capture contextual relationships and used TF-IDF vectorization to convert text into numerical features. Naive Bayes outperformed other models, achieving an F1 score of 0.72, precision of 0.72, and recall of 0.73. It effectively leveraged TF-IDF's sparse representation and class independence assumptions to handle the high-dimensional bi-gram-based feature space. KNN performed the worst (F1: 0.62, precision: 0.53), likely due to difficulties in handling high-dimensional representations [14].

This analysis highlights Naive Bayes' advantage in sparse text classification and the limitations of more complex models on this dataset (see Table 2).

Table 2: Performance of Freely Available Models on the Concerning vs. Non-Concerning Citation Classification Task.

Conventional Machine Learning Models					**Decoder-based Models**				
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Model**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
SVM	0.70	0.68	0.70	0.68	LLaMA 2 7B	0.60	0.21	0.17	0.19
LR	0.69	0.67	0.69	0.66	LLaMA 2 7B Chat	0.59	0.20	0.17	0.18
Decision Tree	0.66	0.63	0.64	0.64	LLaMA 13B	0.75	0.63	0.19	0.29
Random Forest	0.69	0.68	0.69	0.64	LLaMA 13B Chat	0.77	0.59	0.46	0.52
KNN	0.73	0.53	0.73	0.62	LLaMA 2 70B	0.64	0.42	0.82	0.55
Naive Bayes	0.73	0.72	0.73	0.72	LLaMA 2 70B Chat	0.59	0.20	0.17	0.18
XGBoost	0.64	0.62	0.64	0.63	GPT-3.5	0.69	0.45	0.64	0.53
LightGBM	0.64	0.61	0.65	0.62	GPT-4	0.74	0.51	0.54	0.53
LDA	0.71	0.66	0.71	0.66	GPT-4_o	0.77	0.65	0.39	0.48
QDA	0.66	0.65	0.66	0.65	GPT-4_omini	0.76	0.59	0.44	0.51
Sequence-to-Sequence Models					**Encoder-based Models**				
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Model**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
BART-large	0.65	0.65	0.65	0.65	BERT-base	0.87	0.78	0.92	0.85
T5-base	0.73	0.53	0.73	0.62	RoBERTa	0.73	0.53	0.73	0.62
T5-large	0.73	0.53	0.73	0.62	DistilBERT	0.75	0.75	0.75	0.75
Flan-T5-base	0.73	0.53	0.73	0.62	ALBERT	0.74	0.74	0.74	0.74
Flan-T5-large	0.73	0.53	0.73	0.61	Electra	0.77	0.77	0.77	0.77

Encoder-Based LLMs Encoder-based models, such as BERT, are well-suited for classification tasks as they capture bidirectional context through a robust attention mechanism. We evaluated BERT-base, RoBERTa, DistilBERT, ALBERT, and ELECTRA. Each model was subsequently fine-tuned on our dataset. BERT-base yielded the best results, achieving an F1 score of 0.85, a recall of 0.92, and a precision of 0.78. The high recall indicates strong identification of true positives, making this model the most reliable for distinguishing concerning citations with minimal false positives. RoBERTa, ALBERT, DistilBERT, and ELECTRA all showed moderate effectiveness, with F1 scores of 0.73, 0.74, 0.75, and 0.77, respectively, offering balanced but less robust performance than BERT-base. Overall, these findings make the BERT-base model most suitable for this task (see Table 2).

Sequence to Sequence based Models We evaluated BART, T5, and Flan-T5 to assess whether their dual architecture enhances performance in identifying concerning citations. These models process input sequences with an encoder, which creates a context vector, and a decoder that generates a target sequence or class labels for classification tasks. Among these, BART-large achieved the best overall performance, with accuracy, precision, recall, and F1 score all at 0.65. However, its moderate results suggest a limited ability to minimize false positives. The T5 and Flan-T5 models showed better recall (0.73), indicating improved true positive detection, but at the cost of reduced precision (0.53), leading to more false positives. Neither model's scale (base vs. large) nor instruction fine-tuning in Flan-T5 improved performance on any metric. This analysis suggests that while BART performs best, T5 and Flan-T5 models offer a tradeoff between recall and precision.

Decoder-Based LLMs Decoder-based models use a unidirectional structure to generate coherent text by predicting the next token in a sequence. Addition-

ally, large decoder-based models can perform few-shot and zero-shot learning through in-context examples, allowing them to adapt to tasks without retraining. To evaluate if these capabilities apply to our task, we tested the latest Llama2 and ChatGPT models, including their instruction-tuned versions across different scales. We initially framed citation detection as a one-shot classification task [20], providing the model with a task description and a single example in the prompt to establish baseline performance with minimal guidance. To further assess model capabilities, we conducted a few-shot classification experiment [8]. For both setups, we used demonstration-based prompting [4], embedding task-specific examples in the prompt to guide responses. Our results show that the Llama2 70B Base model achieved the highest performance, with an F1 score of 0.55 and a notably high recall of 0.82, indicating strong sensitivity in detecting concerning citations but also a higher rate of false positives. ChatGPT models performed poorly, with F1 scores similar to Llama2 70B. This suggests that detecting concerning citations is a complex task that may not be effectively addressed by training large models solely on general language modeling objectives.

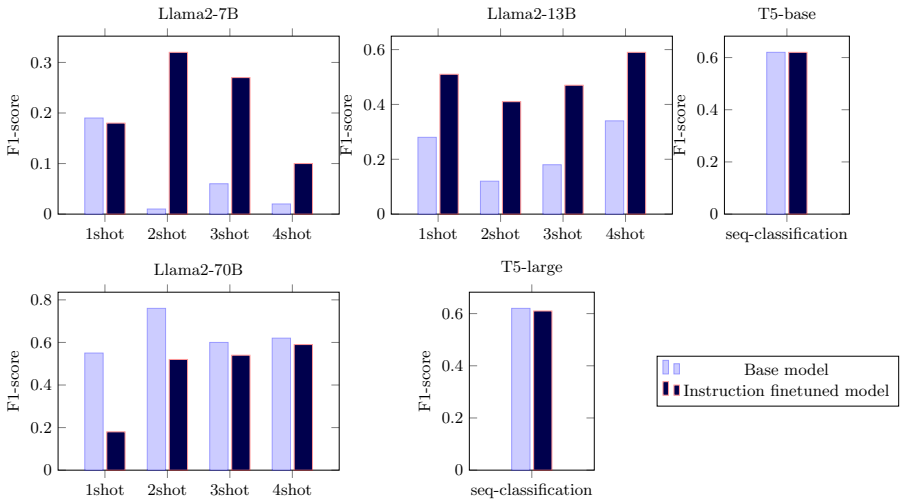

Fig. 2: Comparison of base decoder model and its instruction finetuned/chat models across several settings.

Does instruction fine-tuning improve model performance? For smaller-scale models, like the 7- and 13-billion parameter versions, instruction-fine-tuned models outperform base models, especially in few-shot settings. However, this trend does not hold for larger models, such as the 70-billion parameter model and T5, suggesting that the benefits of instruction fine-tuning are both scale- and model-specific, without a consistent advantage across all models. The superior performance of smaller instruction-fine-tuned models stems from their relatively low base performance on novel or out-of-distribution tasks. Instruction-tuned models leverage in-context examples more effectively due to training objectives

Model	Accuracy	F1-Score
Llama-2 7B Chat	0.58	0.18
Llama-2 12B Chat	0.76	0.51
Gpt3.5	0.69	0.53
Gpt4	0.74	0.53
Gpt4o-mini	0.77	0.48
Gpt4o	0.75	0.50

(a)

Model	1-shot	2-shot	3-shot	4-shot
Llama2-70B				
Llama2-70B	0.55	0.57	0.60	0.62
Llama2-70B-Chat	0.18	0.51	0.55	0.59
Llama2-7B-Chat	0.18	0.32	0.27	0.10
GPT				
GPT-4	0.52	0.58	0.53	0.55
GPT-4o-mini	0.50	0.62	0.53	0.55
GPT-4o	0.48	0.52	0.47	0.52
GPT-3.5	0.53	0.50	0.49	0.53

(a)

Table 3: (a) Comparison of F1-Score for various models with different in-context examples, and (b) Comparing the effect of model scale on Accuracy and F1.

that emphasize instruction-based learning [22]. In contrast, larger models likely benefit from broader task exposure during training, resulting in base models that perform comparably to their instruction-tuned versions on similar tasks (see Figure 2). This implies that larger models may already carry sufficient context through their diverse training data exposure.

Does scale and in-context learning enhance model performance? In Table 3 (a), we compare model performance with their scaled versions, observing a positive correlation between model scale and improvements in accuracy and F1-score, aligning with the general trend in large language models. For the GPT family, performance gains from scaling are moderate, while Llama2 shows substantial improvements with increased model size. For in-context learning, we evaluated up to 4-shot learning, where each model received four example pairs before classification. As shown in Table 3 (b), only the largest Llama model (70 billion parameters) benefits significantly from additional in-context examples; smaller Llama models show minimal performance changes as the number of examples increases. For the ChatGPT models, in-context learning has little effect overall, with two-shot settings often being optimal. These findings suggest that adding more few-shot examples may not effectively boost performance. Instead, expanding the dataset, particularly by including more examples of concerning citations, could be a more impactful approach to improving model performance.

5 Conclusion

This study highlights the significance of citations as a quality control mechanism, demonstrating how early warnings about potentially unreliable research could help prevent the perpetuation of misinformation over an extended period. A retrospective analysis of the data revealed that, in 80% of cases, concerns about later-retracted articles were raised well before the actual retraction occurred. Unfortunately, these early warnings were often overlooked, leading to years of delay in retracting these articles. Taking prompt action on such early warnings could have significantly reduced the long-term impact on scientific integrity. Moreover, we assessed the efficacy of state-of-the-art AI models in identifying citations that

raise concerns about the reliability of cited work. This initial but comprehensive evaluation shows that differentiating between concerning and non-concerning citations remains a challenging task, suggesting the need for further investigation. However, our findings offer valuable early warning indicators of potential retractions, providing a tool for flagging potentially unreliable research. This approach enables researchers and institutions to take steps for follow-up validation, ensuring the integrity of the scientific record and addressing the consequences of retractions in a timely manner. The data and code used in this study are publicly available for further analysis[2].

6 Limitation and Future Work

As a preliminary study, the number of concerning citations—though significant—remains limited in scope. In future research, we plan to incorporate large language models to enhance inference capabilities. Our goal is to develop a more nuanced, evidence-based framework for assessing the plausibility of retractions, enabling a comprehensive evaluation of scientific databases that considers both individual articles and their broader contextual implications within the scientific landscape.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. Gepris 267140244 for the PubPharm – Specialized Information Service for Pharmacy.

References

1. Bolland, M.J., Grey, A., Avenell, A.: Citation of retracted publications: A challenging problem. Accountability in Research **29**(1), 18–25 (2022). https://doi.org/10.1080/08989621.2021.1886933
2. Bordignon, F.: Critical citations in knowledge construction and citation analysis: from paradox to definition. Scientometrics **127**(2), 959–972 (2022). https://doi.org/10.1007/s11192-021-04226-0
3. Bordignon, F., Gambette, P.: A corpus of critical citations contexts. Journal of Open Humanities Data **10** (2024). https://doi.org/10.5334/johd.215
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS 2020) (2020), `https://arxiv.org/abs/2005.14165`
5. Chan, T.H., Kwan, B.S.: What do researchers cite in their literature review sections? an exploratory study of citations in information systems research articles. ibérica (44), 49–74 (2022). https://doi.org/10.17398/2340-2784.44.49
6. Erler, J.T., Bennewith, K.L., Nicolau, M., Dornhöfer, N., Kong, C., Le, Q.T., Chi, J.T.A., Jeffrey, S.S., Giaccia, A.J.: Lysyl oxidase is essential for hypoxia-induced metastasis. Nature **440**(7088), 1222–1226 (2006). https://doi.org/10.1038/nature04695

[2] https://github.com/DASFAA25/Toward-Scientific-Integrity

7. Erler, J.T., Bennewith, K.L., Nicolau, M., Dornhöfer, N., Kong, C., Le, Q.T., Chi, J.T.A., Jeffrey, S.S., Giaccia, A.J.: Retraction note: Lysyl oxidase is essential for hypoxia-induced metastasis (2020). https://doi.org/10.1038/s41586-020-2112-4
8. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better fewshot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021) (2021), https://arxiv.org/abs/2009.07118
9. Gholampour, B., Gholampour, S., Noruzi, A., Arsenault, C., Haertlé, T., Saboury, A.A.: Retracted articles in oncology in the last three decades: frequency, reasons, and themes. Scientometrics **127**(4), 1841–1865 (2022). https://doi.org/10.1007/s11192-022-04305-w
10. Hsiao, T.K., Torvik, V.I.: Opcitance: Citation contexts identified from the pubmed central open access articles. Scientific Data **10**(1), 243 (2023). https://doi.org/10.1038/s41597-023-02134-x
11. Ihsan, I., Qadir, M.A.: Ccro: Citation's context & reasons ontology. IEEE access **7**, 30423–30436 (2019). https://doi.org/DOI: 10.1109/ACCESS.2019.2903450
12. Jiang, X., Cai, C., Fan, W., Liu, T., Chen, J.: Contextualised modelling for effective citation function classification. In: Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval. pp. 93–103 (2022). https://doi.org/10.1145/3582768.358276
13. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the national cancer institute **22**(4), 719–748 (1959). https://doi.org/10.1093/jnci/22.4.719
14. Pestov, V.: Is the k-nn classifier in high dimensions affected by the curse of dimensionality? Computers & Mathematics with Applications **65**(10), 1427–1437 (2013). https://doi.org/doi.org/10.1016/j.camwa.2012.09.011
15. Teixeira da Silva, J.A., Bornemann-Cimenti, H.: Why do some retracted papers continue to be cited? Scientometrics **110**, 365–370 (2017). https://doi.org/10.1007/s11192-016-2178-9
16. Te, S., Barhoumi, A., Lentschat, M., Bordignon, F., Labbé, C., Portet, F.: Citation context classification: Critical vs non-critical. In: Proceedings of the Third Workshop on Scholarly Document Processing. pp. 49–53 (2022)
17. Usman, M., Balke, W.T.: On retraction cascade? citation intention analysis as a quality control mechanism in digital libraries. In: International Conference on Theory and Practice of Digital Libraries. pp. 117–131. Springer (2023). https://doi.org/10.1007/978-3-031-43849-3_11
18. Valenzuela, M., Ha, V., Etzioni, O.: Identifying meaningful citations. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence (2015)
19. Van Noorden, R.: More than 10,000 research papers were retracted in 2023—a new record. Nature **624**(7992), 479–481 (2023). https://doi.org/10.1038/d41586-023-03974-8
20. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., Kavukcuoglu, K.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems (NeurIPS 2016) (2016), https://arxiv.org/abs/1606.04080
21. Xu, X.H., Huang, X.W., Qun, L., Li, Y.N., Wang, Y., Liu, C., Ma, Y., Liu, Q.M., Sun, K., Qian, F., et al.: Two functional loci in the promoter of epas1 gene involved in high-altitude adaptation of tibetans. Scientific reports **4**(1), 7465 (2014). https://doi.org/10.1038/srep07465
22. Zhu, W., Tan, M.: SPT: Learning to selectively insert prompts for better prompt tuning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 11862–11878. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.727