Pub-Guard-LLM: Detecting Retracted Biomedical Articles with Reliable Explanations

Lihu Chen[1], **Shuojie Fu**[1], **Gabriel Freedman**[1], **Cemre Zor**[2], **Guy Martin**[3], **James Kinross**[1], **Uddhav Vaghela**[1], **Ovidiu Serban**[1], **Francesca Toni**[1]
[1] Imperial College London, UK
[2] Amazon Web Services, UK
[3] National Health Service, UK

Abstract

A significant and growing number of published scientific articles ends up being retracted. Many of these retracted articles continue to be cited and influence research or clinical decisions, posing serious societal threats. In this paper, we propose Pub-Guard-LLM, the first large language model-based system tailored to retraction detection for biomedical scientific articles. We provide three application modes for deploying Pub-Guard-LLM: Vanilla Reasoning, Retrieval-Augmented Generation, and Debate, by allowing for textual explanations of prediction in each mode. To assess the performance of our system, we introduce an open-source benchmark, *PubMed Retraction*, comprising over 11K real-world biomedical articles, including metadata and retraction labels. We show that across all modes, Pub-Guard-LLM consistently surpasses the performance of various baselines and provides more *reliable* explanations, namely explanations which are deemed more *relevant* and *coherent* than those generated by the baselines when evaluated by multiple assessment methods. Pub-Guard-LLM can be used to flag potential retraction before peer review. By enhancing both detection performance and explainability in scientific retraction detection, it can contribute to reducing review workloads and preventing the spread of misinformation. The code is available at https://github.com/tigerchen52/pub_guard_llm

1 Introduction

A considerable proportion of published scientific articles are retracted. It is estimated that between 1.5 and 2% of all scientific papers published in 2022 closely resemble content from paper-mill products and retracted articles (Noorden, 2023). In 2023, about 8,000 papers were retracted from Hindawi journals due to their fraudulent origins in paper mills (Parker et al., 2024). Even more concerning, generative AI tools including those built on large language models (LLMs) can produce highly convincing articles that have the potential to bypass existing detection mechanisms (Májovskỳ et al., 2023; Kendall and Teixeira da Silva, 2024), which increases the risk of future retractions. This indicates that the area of problematic publications is constantly adapting and learning to circumvent existing detectors (Májovskỳ et al., 2023; Perkins, 2023).

Retractions of *biomedical* publications are more pronounced than in other fields (Grieneisen and Zhang, 2012; Bik et al., 2016), with a recent finding revealing that over a fifth of newly published medical articles have issues that may lead to retraction (Sabel et al., 2023). Moreover, the problem of retracted articles is not confined to a specific set of offending author institutions. In fact, recent high-profile cases have involved retractions from reputable organizations such as Harvard Medical School (Johnson, 2024). Many retractions happen post-publication, often after flawed papers have already influenced research or clinical decisions. Even after retraction, many articles continue to be cited without acknowledgment of their retraction (Vuong, 2020). This widespread trend poses a serious threat to public health, as misleading medical research can directly influence clinical decisions, leading to ineffective or even harmful treatments for patients. Therefore, it is necessary to develop a detection tool for retraction risk before peer review, which aims to flag articles that contain methodological flaws, metadata inconsistencies, or other issues of scientific integrity. Identifying these potential retractions can help stop flawed manuscripts early, ease reviewer workloads, and prevent the spread of misinformation.

Prior attempts addressing this challenge mainly rely on relatively primitive heuristic techniques (Parker et al., 2022; Shepperd and Yousefi, 2023; Feng et al., 2024), including some based on argument quality (Freedman and Toni, 2024).

While the threat of retracted articles is likely to grow in the future, the AI community has yet to give this issue the attention it deserves. This lack of focus is partly due to two key challenges. First, there is no standard benchmark dataset for evaluating and comparing various retraction detection systems. Second, there are no open-source systems leveraging on LLMs specially designed for this task.

To address these challenges, we introduce *PubMed Retraction*, the first large-scale open-source benchmark for retraction detection in biomedical research, including over 11K *real-world* articles. Our benchmark is designed to consider the diversity across various types of retracted articles. The main goal of this benchmark is to flag potentially retracted articles with minimum information, such as abstracts and metadata,[1] which can help editors and conference organizers to flag potential retractions. Also, we release the first LLM-based system, `Pub-Guard-LLM`, dedicated to the task of retraction detection. `Pub-Guard-LLM` can be deployed in three application modes (see Figure 1): Vanilla Reasoning, Retrieval-Augmented generation, and Debate, each making use of external knowledge and fine-tuning and each returning (textual) explanations of predictions. We show that `Pub-Guard-LLM` can significantly outperform baselines while generating more *reliable* explanations, by being more *relevant* to the explanandum (Kotonya and Toni, 2024) and more *coherent* (Kotonya and Toni, 2020). The three application modes provide options for users to focus on depending on their needs, balancing key factors such as precision and recall of predictions, reliability of explanations, and inference speed. This adaptability makes `Pub-Guard-LLM` highly versatile, accommodating a wide range of use cases in retracted article detection.

With the introduction of the first publicly available benchmark and an open-source LLM specifically designed for retracted article detection, our work aims to serve as a foundational step for future research in this field. However, while technological solutions like ours can help address this issue, we believe the root of the problem lies within academic evaluation systems themselves. Researchers face immense pressure to publish to secure funding and career advancements. As long as these systems prioritize metrics and publication counts over genuine scientific contributions, the arms race between retractions and detectors will persist.

[1]Complex features, such as tables and images, are not the focus of the current version of our benchmark

2 Related Work

The increasing number of scientific retracted papers presents a serious challenge to both the research community and society at large. Previous studies address this challenge by primarily focusing on rule-based heuristics, e.g. author affiliations (Sabel et al., 2023), citation patterns (Shepperd and Yousefi, 2023; Feng et al., 2024), journal impact factors and author networks (Pérez-Neri et al., 2022). Recently, traditional machine learning (Dadkhah et al., 2023; Cabanac et al., 2022) as well as LLM-based models (Freedman and Toni, 2024; Fletcher and Stevenson, 2025) have been proposed to help with retracted article detection.

However, the rapid growth in the number of retracted articles has not been matched by dedicated attention within our field (Byrne et al., 2024). Currently, there is an urgent need for more research into two key aspects. (1) *Benchmarks*. Existing studies generally employ Retraction Watch (The Center for Scientific Integrity, 2018), a blog that monitors and reports on retractions of scientific papers, to analyze retraction behaviors (Shepperd and Yousefi, 2023; Byrne et al., 2022). The Retraction Watch provides a collection of retracted records, but it does not constitute a benchmark on its own for comparing existing methods. There is currently no standardized benchmark including diverse real-world retracted articles. (2) *Open-Source Tools*. While some systems for detecting retracted articles exist, they are often proprietary and controlled by commercial entities that are reluctant to share their technologies (Christopher, 2021). Even though some LLM-based applications have been developed to address news misinformation (Cao et al., 2024) and plagiarism (Wahle et al., 2022), there is a need for open-source LLMs specifically tailored to scientific retraction detection. Such models would enhance transparency and foster collaboration in the fight against problematic articles.

3 *PubMed Retraction* Benchmark

In this work, we aim to detect *retraction risk*, which can be defined as the identification and public flagging of a scholarly paper that contains substantial methodological flaws, metadata inconsistencies, or other issues of scientific integrity, which, if sub-

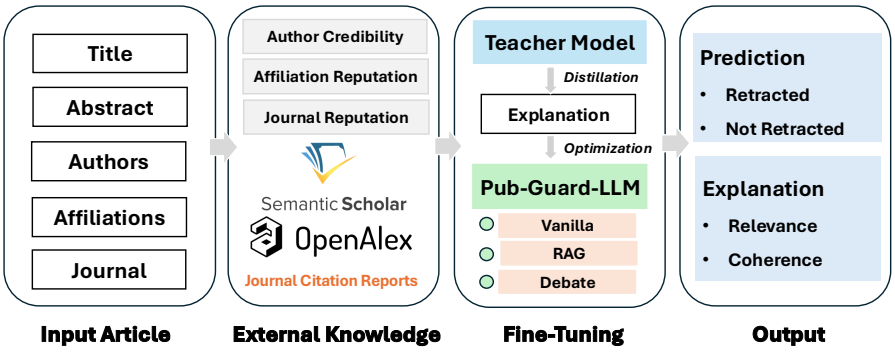


Figure 1: Workflow of the proposed Pub-Guard-LLM (Details in Section 4)

Dataset	General			Cancer			
	Train	Validation	Test	Breast	Lung	Ovarian	Colorectal
Sample	10,000	500	500	300	300	292	300
Retraction Rate	24.5%	25.6%	22.4%	25.0%	38.7%	38.0%	33.0%
High Profile Rate	34.0%	35.9%	33.9%	38.7%	22.4%	29.7%	28.3%

Table 1: Basic statistics of *Pubmed Retraction*. High Profile Rate refers to the articles deemed retracted despite their meta-data being ranked high (details in Appendix A).

stantiated, are likely to lead to retraction.

To address the lack of a standardized benchmark for evaluating and comparing retracted article detection methods, we curate a large open-source benchmark for biomedical articles, *PubMed Retraction*, comprising over 11K real-world biomedical articles, including both non-retracted and retracted publications. We select the biomedical domain due to the higher prevalence of retracted publications therein than in other fields (Grieneisen and Zhang, 2012; Bik et al., 2016), as well as the significant impact of this type of retraction in patient safety, clinical decision-making, and public trust in medical research.

To construct *PubMed Retraction*, we use the PubMed database (Wheeler et al., 2007) and retrieve articles using the Unix command-line tool EDirect (Kans, 2024). To ensure relevance and avoid outdated data, we focus on articles published between 2000 and 2025. Our goal is to promote the use of minimal textual information for retraction detection, so we design the benchmark to include only title, abstract, and metadata. Our benchmark is intended to serve as an early warning system for decision-makers. For example, it can be used before peer review when reviews and citation data are not yet available, which enables journal editors and conference organizers to receive alerts before review assignments are made. Although we fully acknowledge that including full texts, figures, and tables could enhance this task, we aim to develop a simple yet effective solution based on minimal information (title, abstract, metadata) in this work.

Specifically, each article in *PubMed Retraction* is structured to include seven key attributes: *Pubmed ID, Title, Abstract, Authors, Affiliations, Journal, Is Retracted*, with the latter a binary indicator (Yes or No). An article is labeled as *retracted* if the keyword "retracted" appears in its publication type attribute. To maintain a realistic distribution, we include a lower number of retracted articles than legitimate ones, which is aligned with real-world cases. Overall, the benchmark consists of 11,192 articles.

To evaluate generalization capabilities of models, we partition articles based on keywords to create two subsets as described below. This partitioning strategy is designed for a fair assessment of a model's ability to handle out-of-distribution samples, a critical aspect for real-world applications where models encounter unseen data. (1) *General Set*. This set comprises articles covering various diseases and is used for model fine-tuning. It is further split into Train, Validation, and Test subsets. (2) *Cancer Set*. This set consists exclusively of articles related to four types of cancer: Breast, Lung, Ovarian, and Colorectal. Articles of these four categories are excluded from the General Set, which means that these cancers are unseen diseases during fine-tuning and validation. Additionally, our benchmark includes high-profile retracted cases, which makes this task particularly challenging (details in Appendix A).

Formally, ignoring the partitions, *PubMed Re-*

traction can be represented by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each x_i is an article (title, abstract and metadata) and y_i is the binary label. The basic statistics of *Pubmed Retraction* are given in Table 1.

4 Pub-Guard-LLM

Pub-Guard-LLM is an end-to-end system for retracted article detection. Figure 1 provides an overview. First, the input data is enriched with external knowledge. Second, explanations to support article labels (*retracted* or *not*) are obtained by distilling a strong LLM (*Teacher Model*) via zero-shot prompting. Finally, using the article attributes and external knowledge as inputs, and distilled explanations with article labels as outputs, we fine-tune an LLM to create Pub-Guard-LLM. We propose three application modes for Pub-Guard-LLM, resulting in different performances as concerns F1, recall, reliability of explanations, and inference speed.

4.1 External Knowledge Workflow

To complement the seven input attributes (see Section 3) and incorporate task-specific knowledge, we augment external knowledge into Pub-Guard-LLM, improving LLM reasoning—a widely adopted strategy in NLP (Pan et al., 2024). For this, we identify three fundamental factors to assess an article's legitimacy: (1) author credibility, (2) affiliation reputation, and (3) journal reputation, and use the relevant external knowledge to enrich the input.

Author Credibility. In order to evaluate the credibility of authors, we use the Semantic Scholar database (Lo et al., 2019), which contains millions of articles with extensive metadata and citation information, as an external knowledge source. To quantify author reputation, we use the h-index. Since LLMs are not inherently sensitive to numerical values, we map h-index scores into five reputation levels (*very low, low, medium, high, very high*) using a predefined piecewise function (see Appendix A), e.g., *Geoffrey Hinton (author h-index: 188, very high).*

Affiliation Reputation. As an indicator of the reputation of an author's affiliation, we use the average citation count per institution, determined with OpenAlex (Priem et al., 2022) as external knowledge source. This is a fully open index of scholarly works, which provides comprehensive data on the total publications and citations of institutions. The calculated average citation count is then mapped into five reputation levels (*very low, low, medium, high, very high*), e.g., *Harvard University (institution average citation: 64, very high).*

Journal Reputation. We use the ImpactFactor library[2] to obtain the Journal Citation Reports (JCR) partition, which serves as an indicator of a journal's reputation. To enhance interpretability, we assign a human-readable categorized label (*low, medium, high* or *very high*) to each journal based on its JCR ranking, e.g., *Nature (journal JCR: Q1, very high).*

The knowledge retrieved from external resources is integrated into the benchmark $\mathcal{D}$ for reducing duplicated queries, formally resulting in $\mathcal{D}' = \{(x_i, k_i, y_i)\}_{i=1}^{N}$ where k_i is the external information for $(x_i, y_i) \in \mathcal{D}$. Note that, if the inquired knowledge is missing from the external knowledge sources, we use, as k_i, the keyword *"null"* to indicate the absence. We choose to do so rather than omitting the field entirely, as we found it does generally indicate a lower article quality and improves the downstream performance.

4.2 Explanation Distillation

The *PubMed Retraction* benchmark provides only a binary label (*retracted* or *not*) as a target. To construct a model that classifies articles accurately while offering reliable explanations, it is essential to incorporate either human-annotated or machine-generated explanations into the ground truth. Given the high cost of human annotation, we opt for the latter as a more scalable and efficient alternative.

Recent research shows that distilling reasoning capabilities from larger (language) models can significantly enhance the performance of smaller models with fewer learning examples (Li et al., 2022; Hsieh et al., 2023; Shridhar et al., 2023): teacher model rationales provide richer insights into why an input is mapped to a specific output label, capturing task-relevant knowledge missing from the input. What is more interesting is that even if some intermediate steps of the distilled explanations contain inaccuracies, student models can still have enhancements by learning reasoning flows (Li et al., 2023; Hsieh et al., 2023). We use a teacher model to generate explanations for each article via prompting, resulting in $\mathcal{D}'' = \{(x_i, k_i, e_i, y_i)\}_{i=1}^{N}$ where e_i is the corresponding distilled explanations with regard to $(x_i, k_i, y_i) \in \mathcal{D}'$.

Note that instruction-aligned LLMs tend to classify all articles as not retracted, likely to avoid

[2] impact-factor.readthedocs.io/en/latest/

controversial responses. This behavior arises because many LLMs, when fine-tuned for alignment and safety, may become overly conservative and ignore contextually relevant information (Röttger et al., 2024). To mitigate this issue, it is crucial to assign a clear stance to the teacher model. For example, when evaluating a retracted article, we include the prompt: *"You are tasked with providing explanations and making a firm case for why it has issues and should be retracted. You should have a clear stance"*. This ensures strong, rationale-driven explanations instead of the default neutral or overly cautious responses (see Table A3 in Appendix for an example prompt). Additionally, we conduct human evaluations to validate the coherence and relevance of distilled explanations (see Section C.2 in Appendix).

4.3 Fine-Tuning

We leverage the samples in *PubMed Retraction* augmented with the external information and distilled explanations ($\mathcal{D}''$) to fine-tune `Pub-Guard-LLM`. To encourage `Pub-Guard-LLM` to learn both labels and explanations, we introduce a multi-task learning objective. Suppose we have a base model f_θ able to generate label y and explanation e given article x with external information k , i.e., $f_\theta(x, k) = (y, e)$. Then, our goal is to optimize θ for this task.

The model is fine-tuned on y to understand how to identify retracted articles via a binary classification task with:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N} \ell\big(f_\theta(x_i, k_i), y_i\big) \quad (1)$$

where ℓ is the cross entropy between the ground truth and the predicted label. To encourage the model to output explanations behind the decision, we introduce a next-token prediction task with:

$$\mathcal{L}_{\text{explanation}} = \frac{1}{N}\sum_{i=1}^{N} \ell\big(f_\theta(x_i, k_i), e_i\big) \quad (2)$$

where ℓ is the averaged cross entropy between the generated tokens and the distilled explanation from the teacher model.

During fine-tuning, the model learns to predict both label y and explanation e given an article:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{explanation}} \quad (3)$$

where λ is a hyper-parameter that controls the balance between the two objectives. This learning process helps `Pub-Guard-LLM` acquire knowledge beyond simple binary labels, which results in a strong performance with well-reasoned explanations.

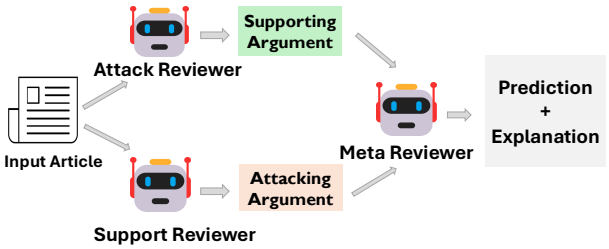


Figure 2: Debate application mode for `Pub-Guard-LLM`

4.4 Application Modes

After fine-tuning, the model f_θ can be leveraged to identify retracted articles while providing explanations. To showcase its usability, we introduce three distinct application modes (see Appendix B.3 for input-output examples for each mode).

Vanilla Reasoning. This refers to reasoning without requiring additional in-context learning samples (*Zero-Shot Reasoning*), which makes predictions solely based on the article: $f_\theta(x, k)$.

Retrieval-Augmented Generation (RAG). $f_\theta(x, k)$ does not store all articles within its parameters, and, in some cases, an input article's main argument may lack support from existing publications. This suggests that the article's findings could be controversial and potentially retracted. To address this, we retrieve relevant top-l articles $\mathcal{A}$ (*all legitimate*) to the input article x from PubMed and use them to validate the claims made in x by passing them as input into the model: $f_\theta(x, k, \mathcal{A})$ (see details in Appendix B.2).

Debate. Recent research has shown that Multi-Agent Debate can enhance both factual accuracy and explainability in LLMs (Liang et al., 2023; Du et al., 2023; Freedman et al., 2024). The core idea involves multiple agents presenting their arguments, while a judge oversees to reach a final decision.

To generate the debate process in our setting (Figure 2), we fine-tune three specialized `Pub-Guard-LLMs` following the steps in Section 4.3: (1) a support reviewer f^s_θ, trained on legitimate articles, generates supporting arguments arg^+; (2) an attack reviewer f^a_θ, trained on retracted articles, generates attacking arguments

Model	$\|\theta\|$	Test	Breast	Lung	Ovarian	Colorectal	Avg
Fine-Tuning Binary Classifiers							
SciBERT	110M	60.9	56.1	74.0	67.2	71.5	66.0
BioLinkBERT	340M	58.1	57.7	75.4	71.0	72.2	66.9
Gatortron	345M	57.6	58.6	71.2	70.9	67.0	65.1
Llama-3.1	8B	60.9	63.2	76.5	72.3	68.3	68.2
Bio-Llama	8B	56.4	61.7	63.9	66.7	61.6	62.1
BioLinkBERT + PEARL	210M	67.6	69.6	79.0	70.5	69.8	71.3
Gatortron + PEARL	445M	66.3	69.3	78.3	76.2	70.8	72.2
Zero-Shot Reasoning							
Llama-3.1-Instruct	8B	29.8	20.0	46.2	43.3	43.6	36.6
OpenScholar	8B	11.5	18.8	20.3	19.9	14.1	16.9
Bio-Llama	8B	37.7	40.7	56.7	59.3	47.9	48.5
PMC-Llama	13B	36.0	36.4	56.7	23.3	25.8	35.6
Ours							
Pub-Guard-LLM (*Vanilla*)	8B	**70.7**	72.8	79.2	78.3	75.6	75.3
Pub-Guard-LLM (*RAG*)	8B	69.8	**76.1**	**82.3**	**79.5**	**76.4**	**76.8**
Pub-Guard-LLM (*Debate*)	8B	70.1	71.5	81.1	78.0	75.7	75.3

Table 2: Performance in detecting retracted articles. We report average F1 scores over three seeds.

arg^-. Thus, given input article (x, k), the support reviewer f_θ^s consistently provides arguments in favor of the article while the attack reviewer f_θ^a presents counterarguments highlighting potential issues. Following this debate, we introduce (3) a meta-reviewer f_θ^{meta} that is responsible for making the final decision based on the arguments presented by both reviewers. To train the meta-reviewer, we use a teacher model to generate debate-based explanations $\hat{e}$, which are then used to fine-tune f_θ^{meta} by reusing the multi-task learning in Equation 3. Overall, the debate-based detection framework is formulated as: $f_\theta^{meta}(x, k, arg^+, arg^-)$, where the meta-reviewer makes a decision by considering both the supporting and attacking arguments.

5 Experimental Set-Up

5.1 Baselines

We compare our approach to the following:

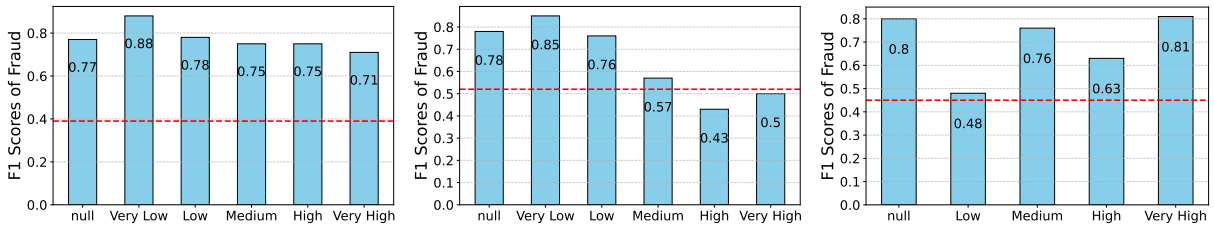
(1) *Fine-tuning Binary Classifiers*. This method appends a classification head—a linear layer—on top of a language model to handle retraction detection. However, this method can only output binary labels (no explanations). The title, abstract, and metadata are concatenated using a special token, [SEP], and this combined textual input is fed into the classifier. Specifically, we use language models: SciBERT (Beltagy et al., 2019), BioLinkBERT (Yasunaga et al., 2022), Gatortron (Yang et al., 2022), Llama-3.1 (Dubey et al., 2024) and Bio-Llama (ContactDoctor, 2024). We also find that introducing additional models to represent metadata separately can be beneficial. Thus, we apply a lightweight short-text model, PEARL (Chen et al., 2024), to represent metadata and obtain embeddings for authors, affiliations, and journal. We continue to use the original language model to represent the title and abstract. Subsequently, these embeddings are concatenated and fed into the classification head. We denote the resulting variant of language model *X* as *"X + PEARL"*.

(2) *Zero-shot Reasoning*. Compared to binary classifiers, LLMs can offer answers with explanations by using prompt-based methods (Kojima et al., 2022). We provide LLMs with a dedicated prompt to obtain binary answers with explanations. Here, we use the following instruction-aligned LLMs: Llama-3.1-Instruct (Dubey et al., 2024), OpenScholar (Asai et al., 2024), Bio-Llama (ContactDoctor, 2024) and PMC-Llama (Wu et al., 2024).

5.2 Implementation Details

All approaches are implemented with PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2020). We use Amazon EC2 *g5.12xlarge* instances with 4×24 GiB for all experiments. The teacher model used for distilling explanations is Mistral-Large (MistralAI, 2024). Our model fine-tunes Llama-3.1-8B (Dubey et al., 2024) using LoRA (Hu et al., 2021) adapters (r=128, lora_alpha=128, lora_dropout=0.1). We train on the train set in the *General* partition for one epoch and AdamW 8-bit optimizer with a learning rate of 1e-4, batch size 4, gradient accumulation of 4 and the λ is 1.

All baseline binary classifiers are fine-tuned with three different seeds: we report their performance based on the best validation set score.



(a) Author Credibility (b) Affiliation Reputation (c) Journal Reputation

Figure 3: Performance on the combined test sets (test plus cancer) of Pub-Guard-LLM (Vanilla) depending on the metadata levels. The horizontal line in each figure represents the F1 score achieved using corresponding heuristic features (see details in Appendix C.1).

6 Results

6.1 Overall Performance

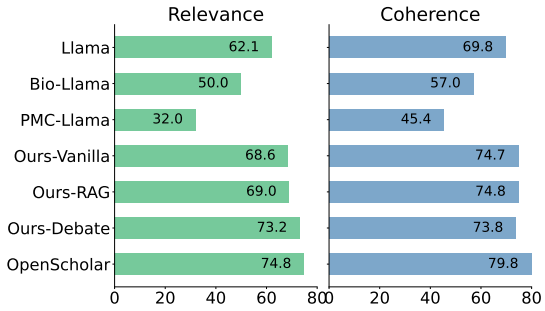
Table 2 gives the results of the comparison of Pub-Guard-LLM with the baselines. First, we note that Pub-Guard-LLM, across all three modes, consistently outperforms the baselines. The performances across the three modes do not exhibit significant differences, but the RAG mode achieves the best score on average, which shows that introducing relevant articles is beneficial to verify the claim quality in abstracts. On the other hand, a key advantage of the debate mode is its notably higher recall, as shown in Appendix Table A1. Achieving high recall is crucial in the retraction detection task, as its primary goal is to flag potential retraction and provide warnings to editors. Second, BERT-based models prove to be cost-effective while achieving competitive performance comparable to larger counterparts in domain-specific tasks, which aligns with prior findings (Lehman et al., 2023; Chen and Varoquaux, 2024). For instance, *Gatortron + PEARL* ranks second overall. Additionally, leveraging top-layer representations of decoder-only models to train classifiers appears suboptimal, as it does not perform better than BERT-based models. Existing LLMs struggle in the zero-shot setting, suggesting that existing LLMs do not have sufficient knowledge for this task. *Note that we do not report the performance of GPT-4o and Mistral-Large because these large instruction-aligned models frequently classify all articles as legitimate to avoid controversial conclusions.*

In summary, all three modes of Pub-Guard-LLM effectively detect retracted articles and each offers distinct advantages. The Vanilla Reasoning mode is efficient and user-friendly, the RAG mode achieves the best overall performance, and the Debate mode excels in recall. This variability makes Pub-Guard-LLM highly versatile, accommodating a wide range of possible user needs in retracted article detection. Additionally, we conduct an ablation study to validate the effect of each component in our system, such as *base LLM*, *Teacher Model*, *External Knowledge*. The results are in Table A2 in the Appendix.

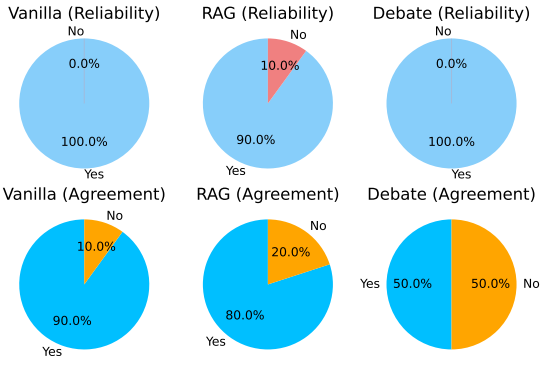
6.2 Pub-Guard-LLM and Metadata Bias

Our proposed method relies on metadata like authors and affiliations. To test whether this metadata introduces biases, specifically against junior researchers or less-known universities, we compare Pub-Guard-LLM (Vanilla) to heuristic methods, which use only individual metadata information to classify retractions and are thus biased. Fo example, we use, as a heuristic method, the credibility level of authors to predict whether a given article should be retracted (see Appendix C.1).

The results are shown in Figure 3. The red horizontal line in each subfigure represents the F1 score achieved by the heuristic methods (i.e. metadata alone). Our findings indicate that Pub-Guard-LLM does not over-rely on a single metadata component. Rather, it integrates both abstracts and three metadata features to make final decisions. This finding proves that Pub-Guard-LLM can mitigate the metadata bias. Furthermore, prestigious researchers, institutions and journals should not be blindly trusted, and these high-profile retractions are harder to detect, as shown by the performance declines on *high* and *very high* groups in Figure 3a and Figure 3b. Note that the *null* is an indicator of low-level credibility and Pub-Guard-LLM is robust against missing information. In summary, high-profile cases pose a significant challenge to retraction detection.



(a) Explanation reliability using LLM-as-a-Judge (average over three runs)



(b) Comparison of explanation reliability and label-explanation agreement in the user study (see details in Appendix C.3)

Figure 4: Explanation evaluation and user study

6.3 Explanation Reliability

6.3.1 Automatic Evaluation

Compared to BERT-based models, a distinct benefit of `Pub-Guard-LLM` is the explanations. To evaluate the reliability of LLM-generated explanations, we consider two dimensions introduced, for different settings, in prior work (Kotonya and Toni, 2020, 2024; Valentino and Freitas, 2024). (1) *Relevance*. A relevant explanation should highlight the particular features and patterns within the article that the model identified as indicative of retraction or not. The model should only reference information that is present in the article, avoiding hallucinations. (2) *Coherence*. This refers to the logical consistency and clarity of the explanation provided by the model for its predictions. Coherence ensures that the explanation flows logically, thereby enhancing the user's understanding of why an article is classified as retracted or not. To measure these properties, we adopt the LLM-as-a-Judge paradigm, which has been shown to align with both controlled and crowdsourced human preferences when using strong LLM judges like GPT-4 (Zheng et al., 2023). We design prompts (see Appendices A5 and A6) for these two dimensions and ask GPT-4o to assess the provided explanations, which assigns a score between 1 and 10. The results, presented in Figure 4a, show that our explanations outperform those of other LLMs, with the exception of OpenScholar. While OpenScholar demonstrates strong explanation quality, its retraction detection capability is significantly lower than other LLMs, with an average F1 score of just 16.9 (see Table 2). Also, note that the debate mode can reduce irrelevant explanations by achieving higher relevance scores than the another two modes. These findings validate that `Pub-Guard-LLM` not only achieves powerful performance but also provides reliable explanations.

6.3.2 User Study

We conducted a pilot user study to evaluate whether the outputs of `Pub-Guard-LLM` align with users' needs. To achieve this, we designed questionnaires covering our three application modes and gathered feedback from three experienced doctors (see details in Appendix C.3). The questionnaire focused on two key aspects: (1) *Reliability:* is the explanation coherent and relevant? (2) *Agreement:* do users agree with the predicted label and the explanation?

Figure 4b presents the user study results. The feedback suggests that doctors find our model's explanations reliable, but there is a discrepancy between expert opinions and machine-generated predictions regarding the debate mode. However, due to the limited number of participants, it is difficult to determine which application mode performs better.

7 Conclusion

In this work, we introduce a publicly available benchmark and the first LLM-based system tailored for retraction detection. Our `Pub-Guard-LLM` not only surpasses existing baselines but also delivers reliable, well-grounded explanations. We aim to provide open-source datasets and LLMs for this task, which helps stop retracted articles early and mitigates misinformation spread.

We emphasize the need for broader efforts to detect retractions in the following key areas. First, high-profile publications should not be blindly

trusted, as retracted publications may still exist within them. Second, our method rely on metadata, which may introduce bias against junior researchers and new institutions. Detecting the above cases requires more than analyzing metadata and abstracts, and future studies should also investigate full texts, tables and images to identify retraction.

Limitations

While our study makes contributions to retracted article detection, we acknowledge several limitations: (1) Catastrophic forgetting. `Pub-Guard-LLM` is a task-specific model that may not perform well in general-purpose reasoning due to catastrophic forgetting (McCloskey and Cohen, 1989)—a phenomenon where a neural model loses previously acquired knowledge upon learning new information. (2) Limited participants in user study. Our user study involved a small number of participants, which limits the ability to systematically compare the three application modes. A more comprehensive study with a larger participant pool is necessary for thorough evaluation. (3) Metadata bias. Our method may create bias against researchers, journals and institutions, as suggested by our user study (Section C.3). (4) Absence of nuance in classification. The current dataset offers only binary labels, *retracted or not*. Future work should refine labels and break down retraction into different reasons, such as scientific misconduct, ethical violations, and honest error.

References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *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 3615–3620.

Elisabeth M Bik, Arturo Casadevall, and Ferric C Fang. 2016. The prevalence of inappropriate image duplication in biomedical research publications. *MBio*, 7(3):10–1128.

Jennifer A Byrne, Anna Abalkina, Olufolake Akinduro-Aje, Jana Christopher, Sarah E Eaton, Nitin Joshi, Ulf Scheffler, Nick H Wise, and Jennifer Wright. 2024. A call for research to address the threat of paper mills. *PLoS Biology*, 22(11):e3002931.

Jennifer A Byrne, Yasunori Park, Reese AK Richardson, Pranujan Pathmendra, Mengyi Sun, and Thomas Stoeger. 2022. Protection of the human gene research literature from contract cheating organizations known as research paper mills. *Nucleic Acids Research*, 50(21):12058–12070.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2022. The'problematic paper screener'automatically selects suspect publications for post-publication (re) assessment. *arXiv e-prints*, pages arXiv–2210.

Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, KP Subbalakshmi, John R Wullert II, Chumki Basu, and David Shallcross. 2024. Can large language models detect misinformation in scientific news reporting? *arXiv preprint arXiv:2402.14268*.

Lihu Chen and Gaël Varoquaux. 2024. What is the role of small models in the llm era: A survey. *arXiv preprint arXiv:2409.06857*.

Lihu Chen, Gael Varoquaux, and Fabian Suchanek. 2024. Learning high-quality and general-purpose phrase representations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 983–994.

Jana Christopher. 2021. The raw truth about paper mills.

ContactDoctor. 2024. Bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B.

Mehdi Dadkhah, Marilyn H Oermann, Mihály Hegedüs, Raghu Raman, and Lóránt Dénes Dávid. 2023. Detection of fake papers in the era of artificial intelligence. *Diagnosis*, 10(4):390–397.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sida Feng, Lingzi Feng, Fang Han, Ye Zhang, Yanqing Ren, Lixue Wang, and Junpeng Yuan. 2024. Citation network analysis of retractions in molecular biology field. *Scientometrics*, 129(8):4795–4817.

Aaron HA Fletcher and Mark Stevenson. 2025. Predicting retracted research: a dataset and machine learning approaches. *Research Integrity and Peer Review*, 10(1):9.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2024. Argumentative large language models for explainable and contestable decision-making. *arXiv preprint arXiv:2405.02079*.

Gabriel Freedman and Francesca Toni. 2024. Detecting scientific fraud using argument mining. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 15–28.

Michael L Grieneisen and Minghua Zhang. 2012. A comprehensive survey of retracted articles from the scholarly literature. *PloS one*, 7(10):e44118.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Carla Johnson. 2024. Harvard medical school affiliate retracts, corrects research. *Fortune*.

Jonathan Kans. 2024. Entrez direct: E-utilities on the unix command line. In *Entrez programming utilities help [Internet]*. National Center for Biotechnology Information (US).

Graham Kendall and Jaime A Teixeira da Silva. 2024. Risks of abuse of large language models, like chatgpt, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, 37(1).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Neema Kotonya and Francesca Toni. 2024. Towards a framework for evaluating explanations in automated fact verification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16364–16377.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, and 1 others. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Martin Májovskỳ, Martin Černỳ, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of medical Internet research*, 25:e46924.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*.

MistralAI. 2024. Mistral large: A general-purpose language model. https://mistral.ai/news/mistral-large-2407/.

Richard Van Noorden. 2023. How big is science's fake-paper problem? *Nature News*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Lisa Parker, Stephanie Boughton, Lisa Bero, and Jennifer A Byrne. 2024. Paper mill challenges: past, present, and future. *Journal of Clinical Epidemiology*, 176:111549.

Lisa Parker, Stephanie Boughton, Rosa Lawrence, and Lisa Bero. 2022. Experts identified warning signs of fraudulent research: a qualitative study to inform a screening tool. *Journal of Clinical Epidemiology*, 151:1–17.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Iván Pérez-Neri, Carlos Pineda, and Hugo Sandoval. 2022. Threats to scholarly research integrity arising from paper mills: a rapid scoping review. *Clinical Rheumatology*, 41(7):2241–2248.

Mike Perkins. 2023. Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond. *Journal of University Teaching and Learning Practice*, 20(2).

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*.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Bernhard A Sabel, Emely Knaack, Gerd Gigerenzer, and Mirela Bilc. 2023. Fake publications in biomedical science: Red-flagging method indicates mass production. *medRxiv*, pages 2023–05.

Martin Shepperd and Leila Yousefi. 2023. An analysis of retracted papers in computer science. *Plos one*, 18(5):e0285383.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

The Center for Scientific Integrity. 2018. The Retraction Watch Database. Accessed: 2025-02-03.

Marco Valentino and André Freitas. 2024. Introductory tutorial: Reasoning with natural language explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 25–31.

Quan Hoang Vuong. 2020. Retractions: The good, the bad, and the ugly. what researchers stand to gain from taking more care to understand errors in the scientific record. *What Researchers Stand to Gain From Taking More Care to Understand Errors in the Scientific Record (February 20, 2020). LSE Impact of Social Sciences (Feb 20, 2020)*.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How large language models are transforming machine-paraphrase plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963.

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, and 1 others. 2007. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*.

C Wu, W Lin, X Zhang, Y Zhang, W Xie, and Y Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association: JAMIA*, pages ocae045–ocae045.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, and 1 others. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Details of Benchmark and External Knowledge

Author Credibility To categorize authors based on their reputation, we define a piecewise function that maps h-index values into discrete reputation levels. This approach ensures that LLMs can interpret author credibility effectively by using textual labels instead of raw numerical values. The categorization is as follows:

- Emerging Researcher (very low): $0 \leq h$-index ≤ 5
- Early Career Researcher (low): $6 \leq h$-index ≤ 15
- Established Researcher (medium): $16 \leq h$-index ≤ 30
- Influential Researcher (high): $31 \leq h$-index ≤ 45
- Leading Expert (very high): h-index > 45

Affiliation Reputation To assess the reputation of an author's affiliation, we use the average citation count per institution, which serves as a key indicator of an institution's scholarly influence. Likewise, we map these values into five distinct reputation levels:

- Developing Institution (very low): $0 \leq$ average citations ≤ 5
- Emerging Institution (low): $6 \leq$ average citations ≤ 15
- Established Institution (medium): $16 \leq$ average citations ≤ 30
- Reputable Institution (high): $31 \leq$ average citations ≤ 45
- World-Class Institution (very high): average citations > 45

Journal Reputation To assess the reputation of a journal, we use the JCR, which classifies journals into four quartiles based on their impact factor. To enhance interpretability, we map JCR quartiles into human-readable reputation levels:

- Top-Level Journal (very high): JCR Q1
- High-Level Journal (high): JCR Q2
- Moderate-Level Journal (medium): JCR Q3
- Low-Level Journal (low): JCR Q4

Definition of High Profile We define an article as high-profile if its metadata includes at least one leading expert (very high) author, or one world-class (very high) affiliation, or is published in a top-level (very high) journal. This definition aligns with our intuition, as articles produced by renowned researchers, prestigious institutions, or highly reputable venues are generally perceived as more trustworthy. We compute the high profile rate of retraction in each subset by: $|\textit{High Profile Articles}|/|\textit{Retracted Articles}|$. These high-profile cases are hard to detect, which makes the task particularly challenging.

Missing Rate In our analysis of the PubMed Retraction dataset, we observed that certain external information may be absent from specific databases. For instance, a journal might not be indexed by the JCR. To quantify this, we computed the missing rates for each feature within our dataset:

- Author Credibility: 10.6% missing
- Affiliation Reputation: 33.4% missing
- Journal Reputation: Journal Indexing: 41.5% missing

B Details of `Pub-Guard-LLM`

B.1 Prompts

Here, we list the prompts used in this work: (1) Explanation Distillation, Table A3; (2) Prompt of Detecting Retracted Articles, Table A4; (3) Prompt of Relevance Evaluation, Table A5; (4) Prompt of Relevance Coherence, Table A6.

B.2 RAG Mode

We retrieve relevant top-l articles $\mathcal{A}$ (*all legitimate*) to the input article x from PubMed and use them to validate the claims made in x. In our experiment, the value l is 5 and the PubMed database is MedRAG PubMed[3] with 3.5M articles. The embedding model here is PubMedBERT[4] and we use cosine similarity to compute the semantic relatedness.

B.3 Output Examples

We provide an input-output example for each application mode: Vanilla Reasoning (Table A7); RAG (Table A8); Debate (Table A9).

[3] https://huggingface.co/datasets/MedRAG/pubmed

[4] https://huggingface.co/NeuML/pubmedbert-base-embeddings

C Details of Experiments

C.1 Heuristic Features

Since we incorporate external knowledge to represent metadata levels, these fine-grained features can serve as indicators for retraction detection. To leverage this, we design three heuristics based on different metadata attributes. The intuition is that articles from prestigious researchers or well-known institutions are generally perceived as more trustworthy. To classify articles as retracted or not, we establish a threshold-based approach using metadata credibility scores. For example, if an article's highest author credibility falls within [*medium, high, very high*], we classify it as legitimate. Author credibility and affiliation reputation are categorized into six levels (including null). We define the three lower levels as indicators of retraction and the three higher levels as indicators of legitimacy. Journal reputation, which consists of five levels, is categorized similarly: the three lower levels indicate potential retraction, while the higher levels suggest legitimacy.

C.2 Evaluation of Distilled Explanation

In this evaluation, we provide our co-authors (NLP people and medics) with questionnaires. This questionnaire is designed to evaluate the quality of explanations generated by an LLM. For each case, the LLM receives information about a medical publication and generates an explanation according to the "Is Retracted" label. This evaluation focuses on this key question " *Is the explanation coherent and relevant? Coherent means the explanation is logical and consistent. Relevant means the explanation is closely connected and appropriate to the provided article information.*" For each article, our co-authors answer one binary question and can add optional comments if they wish. Finally, we evaluate 45 randomly selected articles and 71.1% of these explanations are coherent and relevant.

C.3 User Study

In the user study, we prepared three questionnaires to evaluate the Vanilla Reasoning, RAG and Debate modes independently. Using the Vanilla Reasoning mode, we generated explanations for 300 papers. We randomly selected five retracted papers and five legitimate papers from the result and again randomized the order of the paper to create the questionnaire. For each paper, we added 1) the information of the paper as it was in the model prompt; and 2) the explanations generated by the model. Three questions were asked for each paper: 1) whether the explanation was coherent and relevant; 2) to infer the predicted label from the given explanation; and 3) whether the user agreed with the actual prediction. The same process was followed for the other two questionnaires for the Debate and RAG models, except that, for the Debate model, the information of a supporting viewer and an attacking reviewer was added to the paper description. In total, 30 distinct papers were selected for evaluation by three clinicians.

Given the disagreement between human experts and model-generated outputs regarding the debate mode, we present the first three comments below. We hope these critical insights will be valuable for future studies.

- *I still don't understand what or who the "attacking reviewer" is. Is this a real world person who has completed a peer review? I feel that labelling this as fraudulent without more detailed evidence of actual fraud is not correct. This is very different from this simply being "not very good" science. I don't think it is fair to effectively discredit an author simply because they have a low H factor. Everyone has to start somewhere. Moreover, the interpretation does not provide any evidence at all that the authors have either falsified data, that they have a track record of falsifying data, that the authors are affiliated with authors suspected activity or that they have links to governmental or state actors that may want to push falsified data. There is no comment on funding or conflicts of interest which I would expect to be in this analysis and which is missing. The critique argues that "specific" data is missing from the abstract, but then it does not state what this is and what would be required for the abstract to be robust. In summary, this LLM suggests this is research fraud, when as presented, the data suggest that more simply the quality of the abstract is low. This is an important distinction, that could lead to legal proceedings.*

- *I think this LLM review is a false positive. I think the criticism of the abstract is not founded. They have performed in vivo and in vitro validation studies, and knock out experiments. While detail is missing, this is an*

overly harsh critique of a piece of basic science, that is not pretending to be clinical data. Again, what data is missing form the author's institutions? It doesn't actually state which part of this data is important and why. We know they are from a reputable Japanese university, which seems enough to me. Again, to call in to question the credibility of the scientist is libellous. This would need much stronger evidence of criminal or fraudulent behaviour which is clearly missing. So, I think this is just a case of a not very well written abstract, that has been mis labelled as fraud.

- *The LLM is written in an overly cautious manner, which makes me lack trust in it. Firstly, there is a paradox in its explanation. It states the authors have a high H index, but the institution is emerging... so which are we to judge as being more important? This to me is biasing smaller organisations that may produce outstanding work. We need this model to identify EXCELLENCE in science and to promote it where it exists irrespective of the institution. The factors these models are basing their analysis on e.g. H factor, institutional reputation and author details are not important. I want to understand patterns of publishing behaviours, previous track records of fraud, history of COIs... etc. etc.. and most importantly a robust analysis of the QUALITY of the science. The LLM really provides no insights on this.*

C.4 Ablation Study

To validate the effect of different components in our methodology, we either vary or remove them and observe their impact on model performance, as measured by F1@Val in Table A2. First, we observe that fine-tuning is the most important component, as its removal causes a dramatic performance drop. This indicates it is necessary to adapt LLMs to this task of retraction detection. External knowledge is very beneficial, and the removal leads to a significant decrease. Meanwhile, other components can provide meaningful benefits, which confirms the effectiveness of our methodology.

	Test	Breast	Lung	Ovarian	Colorectal	Avg
Vanilla	84.8	88.5	89.2	85.9	69.7	83.6
RAG	66.9	82.7	87.9	85.6	**81.8**	81.0
Debate	**90.7**	**92.2**	**94.6**	**89.9**	75.9	**88.7**

Table A1: Recall (average over 3 seeds) across modes

	F1@Val	Δ
Pub-Guard-LLM (*Vanilla*)	69.5	-
Varying the Base LLM		
Bio-Llama	68.6	-0.9
Mistral	67.8	-1.7
Varying the Teacher Model		
GPT-4o	68.3	-1.2
Llama-70B	66.4	-3.1
Removing Components		
without Fine-tuning	29.9	-39.6
without Explanation	68.5	-1.0
without External Knowledge	59.8	-9.7

Table A2: Ablation study results

Prompt of Explanation Distillation
This following article is a {label} article.
You are tasked with providing explanations and making a firm case for why it has issues and should be {label}.
This is the given article:
Title: {Title}
Abstract: {Abstract}
Authors: {Authors}
Affiliations: {Affiliations}
Journal: {Journal}
Based on the provided information, give a very short and concise explanation why it should be {label}
Here are some factors to consider:
(1) the reputation of the journal (whether it is a top journal with a rigorous peer review process)
(2) the reputation of the authors and their affiliations (whether the authors have a history of misconduct in their research)
(3) the integrity of the title and abstract (e.g., controversial topic, using made-up data, plagiarism, etc)
In addition, if check if Email addresses are provided, check if it conforms to institutional format. Otherwise, no need to make comments about the Email adresses.
Only provide the key reason briefly (in {Token} words) and do not repeat the title and sentences of abstract.
Note that we use the token 'null' to indicate that additional information is missing from the corresponding databases.
Write a separate sentence for each feature to avoid creating long and complex sentences.

Table A3: Example prompt used for distilling explanations

Prompt of Detecting Retracted Articles

You are tasked with determining whether a given research paper should be retracted.

To make this judgment, analyze the provided title, abstract, author information, institutional affiliation, and publishing journal.

Here are some factors to consider:

(1) the reputation of the journal (whether it is a top journal with a rigorous peer review process)

(2) the reputation of the authors and their affiliations (whether the authors have a history of misconduct in their research)

(3) the integrity of the title and abstract (e.g., controversial topic, using made-up data, plagiarism, etc)

In addition, if check if Email addresses are provided, check if it conforms to institutional format. Otherwise, no need to make comments about the Email adresses.

Use the examples below as guidance:

Example:

Title: `{Title}`

Abstract: `{Abstract}`

Authors: `{Authors}`

Affiliations: `{Affiliations}`

Journal: `{Journal}`

This is the given article:

Title: `{Title}`

Abstract: `{Abstract}`

Authors: `{Authors}`

Affiliations: `{Affiliations}`

Journal: `{Journal}`

Please first provide your prediction of whether this paper should be retracted and then provide your assessment and explanation.

Label (answer Yes or No):

Table A4: The prompt used to query LLMs.

Prompt of Relevance Evaluation

Evaluate the relevance of explanation provided for the model's prediction of the retracted articles.

Consider the following criteria:

Fidelity: Does the explanation highlight specific features and patterns within the article?

Accuracy: Does the explanation avoid mentioning elements that do not exist in the article? The model should not make hallucinated explanations.

Based on these criteria, rate the explanation's relevance on a scale from 1 to 10.

This is the given article:

Title: `{Title}`

Abstract: `{Abstract}`

Authors: `{Authors}`

Affiliations: `{Affiliations}`

Journal: `{Journal}`

This is the prediction and explanation:

Prediction and Explanation: `{explanation}`

Rate the relevance, only provide the score:

Table A5: Example prompt used for evaluating the relevance of explanations

Prompt of Coherence Evaluation

Evaluate the coherence of the explanation provided for the model's prediction of retracted articles. Consider the following criteria:

Coherence: Does the explanation maintain logical consistency and clarity in its reasoning?

Structure: Is the explanation well-organized, with ideas flowing logically from one to another?

Based on these criteria, rate the explanation's coherence on a scale from 1 to 10.

This is the given article:

Title: `{Title}`

Abstract: `{Abstract}`

Authors: `{Authors}`

Affiliations: `{Affiliations}`

Journal: `{Journal}`

This is the prediction and explanation:

Prediction and Explanation: `{explanation}`

Rate the coherence, only provide the score:

Table A6: Example prompt used for evaluating the coherence of explanations

An Example of the Vanilla Reasoning Mode

User Query with External Knowledge:

Title: Changes and Influencing Factors of Cognitive Impairment in Patients with Breast Cancer

Abstract: To investigate the changes in cognitive function and its influencing factors in patients with breast cancer after chemotherapy, to provide a scientific basis for further cognitive correction therapy. In this study, general information on age, marital status, and chemotherapy regimen was collected from 172 breast cancer chemotherapy patients. 172 patients with breast cancer undergoing chemotherapy were investigated by convenience sampling method, and the subjects were tested one-on-one using the Chinese version of the MATRICS Consensus Cognitive Battery (MCCB) computer system.

Authors: Cui (author h-index: 6, Early Career Researcher); Song (author h-index: 1, Emerging Researcher); Zhang (author h-index: 7, Early Career Researcher)

Institutions: College of Nursing, Jinzhou Medical University, Jinzhou, Liaoning 121001, China. (institution average citation: 9.0, Emerging Institution); Department of Intensive Care Medicine, Liaocheng People's Hospital, Liaocheng, Shandong 252000, China. (institution average citation: 10.0, Emerging Institution);

Journal: evidence-based complementary and alternative medicine : ecam (null)

Pub-Guard-LLM:

Prediction:

Yes (*Retracted*)

Explanation:

The article should be retracted due to concerns about the integrity of the study. The authors' affiliations are inconsistent, with some listed multiple times, suggesting a potential error or lack of proper collaboration. The journal in which it was published is not specified, making it difficult to assess the rigor of the peer review process. The study's sample size is relatively small, which may limit the generalizability of the findings. Additionally, the abstract does not provide sufficient detail about the methods used, making it difficult to assess the validity of the results.

Table A7: An input-output example of the Vanilla Reasoning mode

An Example of the RAG Mode

User Query with External Knoledge:

Title: Changes and Influencing Factors of Cognitive Impairment in Patients with Breast Cancer

Abstract: To investigate the changes in cognitive function and its influencing factors in patients with breast cancer after chemotherapy, to provide a scientific basis for further cognitive correction therapy. In this study, general information on age, marital status, and chemotherapy regimen was collected from 172 breast cancer chemotherapy patients. 172 patients with breast cancer undergoing chemotherapy were investigated by convenience sampling method, and the subjects were tested one-on-one using the Chinese version of the MATRICS Consensus Cognitive Battery (MCCB) computer system.

Authors: Cui (author h-index: 6, Early Career Researcher); Song (author h-index: 1, Emerging Researcher); Zhang (author h-index: 7, Early Career Researcher)

Institutions: College of Nursing, Jinzhou Medical University, Jinzhou, Liaoning 121001, China. (institution average citation: 9.0, Emerging Institution); Department of Intensive Care Medicine, Liaocheng People's Hospital, Liaocheng, Shandong 252000, China. (institution average citation: 10.0, Emerging Institution);

Journal: evidence-based complementary and alternative medicine : ecam (null)

RAG Results:

In the meantime, a set of legitimate papers is provided, each containing highly relevant articles. Use these articles as evidence to assess whether the given paper is potentially retracted. Look for indicators such as controversial topics, unsupported claims, or inconsistencies. Below are examples of relevant papers:

Example 1:

Title: `{Title}` Abstract: `{Abstract}`

.....

Pub-Guard-LLM:

Prediction:

Yes (*Retracted*)

Explanation:

The article should be retracted due to the lack of reputation of the journal, as it is not listed in the provided information. Additionally, the authors have relatively low h-index scores, indicating a lack of established reputation in their field. Furthermore, the study's sample size is small, and the results may not be generalizable to a larger population. Lastly, the article lacks information on the study's methodology and statistical analysis, raising concerns about the validity of the findings

Table A8: An input-output example for the RAG mode

An Example of the Debate Mode

User Query with External Knowledge:

Title: Changes and Influencing Factors of Cognitive Impairment in Patients with Breast Cancer

Abstract: To investigate the changes in cognitive function and its influencing factors in patients with breast cancer after chemotherapy, to provide a scientific basis for further cognitive correction therapy. In this study, general information on age, marital status, and chemotherapy regimen was collected from 172 breast cancer chemotherapy patients. 172 patients with breast cancer undergoing chemotherapy were investigated by convenience sampling method, and the subjects were tested one-on-one using the Chinese version of the MATRICS Consensus Cognitive Battery (MCCB) computer system.

Authors: Cui (author h-index: 6, Early Career Researcher); Song (author h-index: 1, Emerging Researcher); Zhang (author h-index: 7, Early Career Researcher)

Institutions: College of Nursing, Jinzhou Medical University, Jinzhou, Liaoning 121001, China. (institution average citation: 9.0, Emerging Institution); Department of Intensive Care Medicine, Liaocheng People's Hospital, Liaocheng, Shandong 252000, China. (institution average citation: 10.0, Emerging Institution);

Journal: evidence-based complementary and alternative medicine : ecam (null)

Supporting Reviewer:

The article appears legitimate based on the following factors:

1. Journal Reputation: The article is published in evidence-based complementary and alternative medicine : ecam, a journal dedicated to complementary and alternative medicine, suggesting a focus on rigorous scientific research in this field. 2. Author Reputation: The authors are affiliated with recognized institutions, and their h-index scores, though relatively low, indicate they have published and been cited in other academic works. 3. Title and Abstract Integrity: The title and abstract present a clear research question, methodology, and findings, with no apparent signs of controversial topics, made-up data, or plagiarism. 4. Institutional Affiliation: The authors are associated with various departments and colleges within recognized universities, suggesting a multidisciplinary approach to the research. 5. Missing Information: The 'null' values indicate missing information, but this does not necessarily delegitimize the article, as it may simply be a database issue.

Attacking Reviewer:

The article should be retracted due to the lack of reputation of the journal, as it is not specified and could indicate a lack of rigorous peer review. The authors have relatively low h-indices, suggesting they are early in their careers or have not yet produced highly impactful research. The study's findings may not be reliable as the sample size is not specified, and the results could be biased due to the convenience sampling method used. Additionally, the absence of information about the Department of Radiation Oncology at Jinzhou Medical University raises concerns about the credibility of the research.

Pub-Guard-LLM:

Prediction:

Yes (*Retracted*)

Explanation:

The article should be retracted due to concerns about the credibility of the research. The journal's reputation is questionable as it is not specified, suggesting a lack of rigorous peer review. The authors' relatively low h-indices and the absence of information about one of the institutions raise questions about their expertise and the validity of the research. Furthermore, the study's findings may not be reliable due to the lack of information about the sample size and the use of convenience sampling, which could introduce bias.

Table A9: An input-output example for the Debate mode